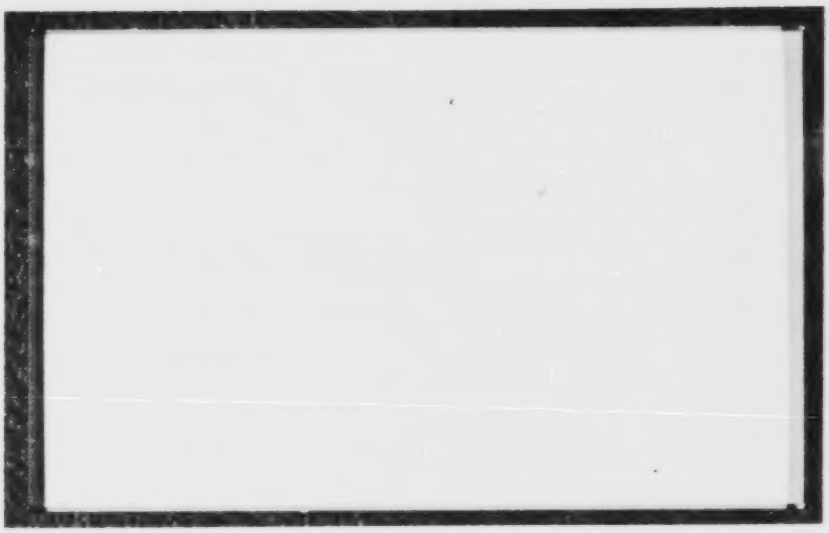
UNIVERSITÉ DE MONTRÉAL



CENTRE DE RECHERCHE SUR LES TRANSPORTS

CASE POSTALE 6128, SUCCURSALE A, MONTRÉAL, QUÉBEC, H3C 3J7

SUR L'ÉTUDE DE FILE D'ATTENTE HYPERCUBE DANS LE CONTEXTE DU DÉPLOIEMENT DES AMBULANCES À MONTRÉAL

par

Jean Choquette

Mémoire de M.Sc., département d'informatique et de recherche opérationnelle, dirigé par Jean-Marc Rousseau et Robert Cléroux

Cette étude a été publiée grâce à une subvention du fonds F.C.A.R. pour l'aide et le soutien à la recherche.

Université de Montréal

Centre de recherche sur les transports - Publication #450

Juillet 1985

Dépôt légal - 2e trimestre 1986

FORMULAIRE DE DOCUMENTATION POUR RAPPORT

1. Titre

Sur l'étude du modèle de file d'attente hypercube dans le contexte du déploiement des ambulances à Montréal

2. Date du rapport

Juillet 1985

3. No de rapport du C.R.T.

450

4. No I.S.B.N.

5. Auteur(s)

Jean Choquette

6. Nom et adresse de l'organisme subventionnaire

7. Résumé : Le présent mémoire traite des problèmes reliés au déploiement d'une flotte d'ambulances sur le territoire de la ville de Montréal (Urgences-Santé). On s'attardera, en particulier, à l'étude d'un modèle de file d'attente (hypercube) pour lequel les serveurs sont répartis dans l'espace. Ce modèle sert à prédire les performances des ambulances (temps moyens de déplacement, charges de travail ...) pour une situation qui est décrite par une série d'hypothèses parfois simplificatrices. Il s'agit de voir si les résultats fournis par le modèle peuvent être utilisés lorsque certaines de ces hypothèses sont violées de façon à se rapprocher d'un contexte réel. Un outil de simulation sera utilisé afin de mettre le modèle à l'épreuve. De plus, une analyse de données sera faite à partir de données provenant d'Urgences Santé. Les buts de cette analyse sont de vérifier les hypothèses statistiques du modèle (arrivée de Poisson pour les appels, temps de service exponentiel) ainsi que d'estimer les différents paramètres (taux d'appels, temps moyen de service ...) que nécessiteront le modèle lui-même et l'outil de simulation.

8. Mots clés :

ambulances; Urgences Santé; localisation; hypercube; simulation.

9. Nombre de pages : 114

10. Code de prix : A05

TABLE DES MATIÈRES

		Page
SOMMAIRE		
CHAPITRE PREMIER - INTRODUCTION		1
1.1	Généralités	1
1.2	Problématique des systèmes d'urgence ambulanciers	2
	1.2.1. Problèmes et sous-problèmes du déploiement	2
	1.2.2 Localisation	2
	1.2.3 Répartition ("Dispatching")	3
	1.2.4 Relocalisation	4
	1.2.5 Problèmes connexes	5
1.3	Situation à Montréal - Urgences Santé	7
	1.3.1 Contexte	7
	1.3.2 Intervenants	7
	1.3.3 Services inter-établissements	8
	1.3.4 Services d'urgence (appels de type 4242)	9
	1.3.5 Particularité de la situation à Montréal	11
1.4	Motivations du projet du C.R.T.	11
1.5	Buts de la présente thèse	12
CHAPITRE DEUXIÈME - MESURES DES PERFORMANCES ET OBJECTIFS		15
2.1	Généralités	15
2.2	Description et justification des mesures utilisées	16
	2.2.1 Utilité des mesures	16
	2.2.2 Temps de réponse	16
	2.2.3 Charges de travail	18
2.3	Objectifs	18
2.4	Aspects aléatoires du système	20
CHAPITRE TROISIÈME - SUR L'ÉTAT DES CONNAISSANCES ACTUELLES DANS LE DOMAINE DES SYSTEMES D'URGENCE		22
3.1	Généralités	22
3.2	Outils d'évaluation des performances du système	23

TABLE DES MATIÈRES (suite)

		Page
3.3	Modèles de recherche de solutions	25
3.3.1	Modèles de localisation	25
3.3.2	Relocalisation	27
3.3.3	Répartition	29
CHAPITRE QUATRIÈME - DESCRIPTION DU MODÈLE HYPERCUBE ET DE LA METHODOLOGIE		31
4.1	Buts	31
4.2	Description du modèle hypercube	33
4.2.1	Idée générale	33
4.2.2	Hypothèses du modèle	33
4.2.3	Formulation mathématique	36
4.2.4	Obtention des équations à l'état d'équilibre et leur solution	39
4.2.5	Robustesse du modèle	40
4.3	Méthodologie	41
CHAPITRE CINQUIÈME - ANALYSE DE DONNÉES		43
5.1	Introduction	43
5.2	Données	44
5.2.1	Informations nécessaires	44
5.2.2	Collecte de données	44
5.2.3	Description des informations recueillies	46
5.2.4	Validité des données	46
5.3	Analyse de la demande	47
5.3.1	Buts	47
5.3.2	Visualisation des données	48
5.3.3	Étude de la demande en fonction du temps	52
5.3.4	Conclusion	59
5.4	Vérification des hypothèses statistiques du modèle hypercube	59
5.4.1	Arrivées des demandes Poisson	59
5.4.2	Temps de service	70
5.5	Temps de réponse et de déplacement	73
5.5.1	Temps de réponse	73
5.5.2	Note sur le temps de déplacement (affectation-arrivée sur les lieux)	79

TABLE DES MATIÈRES (suite)

		Page
CHAPITRE SIXIÈME - RÉSULTATS		81
6.1	Rappels des buts	81
6.2	Programme de simulation	82
	6.2.1 Description	82
	6.2.2 Base commune pour la simulation et le modèle analytique	
6.3	Description des simulations	
	6.3.1 Types de simulation	84
	6.3.2 Génération des appels	86
6.4	Analyse des résultats	87
	6.4.1 Construction d'intervalles de confiance pour les mesures	
	6.4.2 Déterminer le début et la taille des simulations	92
	6.4.3 Présentation des résultats et discussion	92
6.5	Conclusion	97
CHAPITRE SEPTIÈME - CONCLUSION		101

SOMMAIRE

Le présent mémoire traite des problèmes reliés au déploiement d'une flotte d'ambulances sur le territoire de la ville de Montréal (Urgences-Santé). On s'attardera, en particulier, à l'étude d'un modèle de file d'attente (hypercube) pour lequel les serveurs sont répartis dans l'espace. Ce modèle sert à prédire les performances des ambulances (temps moyens de déplacement, charges de travail ...) pour une situation qui est décrite par une série d'hypothèses parfois simplificatrices. Il s'agit de voir si les résultats fournis par le modèle peuvent être utilisés lorsque certaines de ces hypothèses sont violées de façon à se rapprocher d'un contexte réel. Un outil de simulation sera utilisé afin de mettre le modèle à l'épreuve. De plus, une analyse de données sera faite à partir de données provenant d'Urgences Santé. Les buts de cette analyse sont de vérifier les hypothèses statistiques du modèle (arrivée de Poisson pour les appels, temps de service exponentiel) ainsi que d'estimer les différents paramètres (taux d'appels, temps moyen de service ...) que nécessiteront le modèle lui-même et l'outil de simulation.

CHAPITRE I
INTRODUCTION

1.1 Généralités

De nos jours, une grande partie de la population mondiale habite en région urbaine. L'organisation de telles agglomérations exige des structures administratives élaborées. En particulier, la planification et l'exploitation de certains types de service (transport en commun, police, ...) sont devenues de plus en plus complexes. On retrouve donc aujourd'hui des services urbains comme les pompiers, la police, les ambulances, la livraison, les services de réparation à domicile, etc. Ils sont caractérisés par le fait qu'une flotte de véhicules est disséminée sur le territoire en prévision des appels à venir. Lorsqu'un d'eux survient, une ou plusieurs unités de service doit ou doivent se déplacer pour aller satisfaire l'usager ou le client.

On peut séparer cette classe de services en deux catégories : les services d'urgence et les autres. Lorsque l'on considère la notion d'urgence, on doit faire intervenir la rapidité avec laquelle la ou les unités arrivent sur les lieux de l'appel.

Dans ce mémoire, il sera question de services d'urgence et plus particulièrement de services d'urgence ambulanciers. Dans ce domaine, on doit pouvoir répondre à une variété d'appels qui vont des moins graves (transport à l'hôpital de personnes invalides pour y subir des examens par exemple) jusqu'à des cas de personnes grièvement blessées ou malades. En raison des vies humaines qui sont en jeu, les responsabilités sont très grandes.

Dans certaines grandes villes (comme Toronto, Montréal ou Austin, Texas, ...) on a centralisé l'organisation des services ambulanciers plutôt que d'en laisser la responsabilité à des intervenants locaux (centres hospitaliers, compagnies privées). On obtient ainsi une simplification de l'accès des usagers au service par la mise en place d'un numéro de téléphone unique pour y faire appel. De plus, en regroupant les ressources, on peut, à l'aide d'une utilisation rationnelle de celles-ci, offrir un service de qualité à un moindre coût.

La planification et l'exploitation d'un service d'urgence ambulancier centralisé représente toute une tâche à réaliser. L'exercice est rendu d'autant plus difficile si l'on vise une organisation optimale. Nous allons maintenant examiner la problématique qu'on y rencontre.

1.2 Problématique des systèmes d'urgence ambulanciers

1.2.1 Problèmes et sous-problèmes du déploiement

Nous ne nous intéressons pas ici aux décisions médicales liées à l'exploitation d'un système ambulancier. La qualification médicale des ambulanciers, le matériel à bord des ambulances, l'établissement de normes à respecter dans le service (temps de réponse à ne pas excéder par exemple), pour ne citer que celles-ci, ne sont pas au nombre des préoccupations de ce mémoire. Nous allons plutôt nous attarder au problème du déploiement des unités d'urgence sur le territoire. Il peut être divisé en trois sous-problèmes : la localisation, la répartition et la relocalisation. Ces trois sous-problèmes sont évidemment interdépendants.

1.2.2 Localisation

Essentiellement, le problème de la localisation des unités se ramène à déterminer en tout temps

1) Combien de véhicules sont nécessaires pour l'ensemble du territoire de façon à satisfaire des critères de qualité de service préétablis (dont nous parlerons au chapitre 2)?

2) où doit-on placer chacun de ces véhicules pour satisfaire au mieux les mêmes critères?

Manifestement, ces deux questions sont reliées l'une à l'autre. En effet, une mauvaise distribution des véhicules sur le territoire nécessitera davantage de ressources pour atteindre les mêmes critères.

La solution à ce problème dépend de plusieurs paramètres. Entre autres, elle dépend :

- du taux d'appel : plus il est élevé, plus le nombre de véhicules à affecter est grand;
- de la répartition géographique des appels : les régions à forte densité d'appels requerront davantage de ressources;
- du temps de service : plus il est long, plus une unité est occupée longtemps et moins elle est disponible pour répondre à d'autres appels, et par conséquent, plus il faut de ressources.

1.2.3 Répartition ("Dispatching")

Étant donné l'arrivée aléatoire d'une demande en un lieu et à un moment, la répartition consiste à déterminer :

1) le nombre d'unités à envoyer, de même que leur(s) type(s) (lorsqu'on dispose de plusieurs types) en fonction de la gravité et de la nature de l'appel;

2) si les unités doivent être envoyées immédiatement ou non, compte tenu de l'urgence des appels et des véhicules disponibles;

3) quelle(s) unité(s), parmi celle(s) disponible(s), doit(doivent) être désignée(s) pour répondre à l'appel.

On remarquera que les points 1 et 2 font appel à des critères d'ordre médical.

1.2.4 Relocalisation

Il arrive qu'une urgence (comme un grave accident de la route) puisse monopoliser une grande partie des effectifs d'une région, créant un "trou" dans la couverture. Cela peut aussi se produire lorsque survient, dans un même secteur, une série d'appels très rapprochés dans le temps. Il est alors utile de pouvoir déplacer des unités disponibles, localisées dans d'autres régions, pour assurer une meilleure couverture globale, et plus spécifiquement, de la région découverte.

La relocalisation se définit comme le changement des emplacements des véhicules en disponibilité. Le but de toute relocalisation est de se donner la meilleure protection possible contre les événements futurs en modifiant adéquatement la manière dont les unités disponibles sont localisées sur le territoire.

Une relocalisation peut survenir dans plusieurs circonstances. D'une manière générale, il peut être avantageux de faire une relocalisation lorsqu'un véhicule disponible devient occupé suite à un appel, ou inversement, lorsqu'un véhicule occupé devient disponible lors d'une fin de service.

Le problème, dans un contexte où il est possible de relocaliser des unités, se ramène à identifier les moments opportuns où l'on peut relocaliser et la manière dont cela doit être fait. Dans ce dernier cas, on doit identifier quelles unités sont relocalisées, où et comment elles le sont (c'est-à-dire quelles routes empruntent-elles pour atteindre leurs nouveaux emplacements).

Note : Dans plusieurs villes, les autorités ne bénéficient pas d'une telle souplesse. Les unités sont localisées dans des casernes et aucune relocalisation n'est permise.

1.2.5 Problèmes connexes

Sous la considération de critères de qualité de service prédéfinis, on doit déterminer les effectifs à affecter à la couverture en tout temps. Dans l'optique d'une utilisation rationelle des ressources, il importe également de trouver le nombre minimal d'unités à utiliser tout en gardant satisfaits les mêmes critères de qualité de service. La solution à ce problème est intimement liée aux emplacements choisis pour localiser les unités (problème de localisation). Cette solution dépend aussi des politiques de répartition, de même que des politiques de relocalisation des unités (s'il y a lieu).

Le fait de planifier des ressources nécessite des prédictions des taux d'appel pour les différentes périodes de la journée, de la semaine et de l'année. Cela doit être fait à partir de données sur le passé des opérations. Enfin, l'affectation de ressources entraîne avec elle l'affectation de personnel. Un autre problème consiste alors à confectionner l'horaire du personnel de façon à satisfaire les contraintes de demandes des usagers et celles des conventions de travail des ambulanciers.

1.2.6 Complexité de la situation

Considérant d'une part l'aspect aléaloire de l'arrivée des appels tant par le moment que par le lieu, et d'autre part, la nature du service qui induit des temps de service qui sont eux-aussi aléatoires, il découle une situation non déterministe. Il apparaît alors un élément de congestion qui fait qu'on ne peut être certain de la disponibilité d'une unité donnée. On devra prévoir un mécanisme de relève lorsque certaines unités sont occupées. C'est ce contexte qui donne toute sa richesse aux problèmes que nous avons exposés précédemment.

Un autre élément de complexité résulte de la situation qui est continuellement changeante en raison des taux d'appel qui fluctuent durant la journée. Par conséquent, on ne pourra considérer qu'un seul ensemble de solutions pour une situation dont tous les paramètres (taux d'arrivée des appels, temps de service moyen des unités, etc.) sont connus et fixés. (Une telle situation est appelée état d'équilibre). Par exemple, on ne pourra planifier un seul nombre de ressources pour l'ensemble de la journée; on devra tenir compte des différents taux d'appel. On verra donc à affecter plus d'unités pour les périodes de pointe pour offrir un service de qualité. Par contre, il faudra ne pas affecter trop d'unités dans les périodes creuses afin de ne pas utiliser des ressources pour rien.

Jusqu'à maintenant, nous avons parlé d'ambulances. Toutefois, dans un service ambulancier, il peut y avoir plusieurs types de véhicules (ambulances, voitures avec médecins, para-médics). On doit alors se demander comment déployer la flotte des différents types de véhicules de façon à assurer la meilleure coordination entre les différents services qu'elles offrent. Si l'on ajoute à cela le fait que l'on peut associer à chaque appel un niveau de priorité selon la rapidité d'intervention requise, on voit alors que le problème est très complexe.

Enfin, la définition même des critères de qualité de service pose un certain problème. Celle-ci s'exprime en termes d'exigences médicales (on doit traiter les patients le plus adéquatement possible), mais aussi en termes d'équité (on doit être en mesure d'intervenir rapidement auprès de tout usager où qu'il soit). Le problème est de traduire ces critères de qualité en termes quantitatifs (minimiser le temps moyen de réponse, minimiser le temps maximal de réponse, etc.).

Nous nous intéresserons à l'étude du système d'urgence ambulancier à Montréal. Nous avons exposé plus haut les problèmes communs à la grande majorité des systèmes. Cependant, chaque ville fournit ses particularités ajoutant des éléments nouveaux au problème. Nous allons maintenant nous attarder à regarder celles que l'on rencontre à Montréal.

1.3 Situation à Montréal : Urgences Santé

1.3.1 Contexte

L'organisme chargé d'offrir les services d'urgence médicale sur le territoire de l'île de Montréal et de l'île Jésus porte le nom d'Urgences-Santé. C'est un organisme para-gouvernemental faisant partie du Conseil régional des Services Sociaux de Santé du Montréal Métropolitain (CRSSSMM). Il est tenu d'offrir ses services à toute personne qui en fait la demande. Les principaux services consistent

- à transporter des malades et des blessés dans des centres hospitaliers (CH),
- à transporter des malades entre centres hospitaliers,
- à dispenser des soins pré-hospitaliers.

Les gestionnaires d'Urgences Santé coordonnent l'ensemble de ces opérations. En conséquence, ils doivent aussi se charger de l'appareil administratif relié à la planification et à l'exploitation des ressources.

1.3.2 Intervenants

Pour répondre aux différents appels, Urgences Santé dispose de plusieurs ressources. Pour transporter les patients, on peut utiliser une ambulance à une, deux ou trois civières. Les ambulanciers ont la formation pour dispenser certains soins pré-hospitaliers. On peut également compter sur l'envoi de médecins (véhicules-médecins ne faisant pas le transport de patients). Finalement, on peut compter sur la présence de superviseurs.

Ceux-ci ont pour fonction de coordonner les activités dans les cas nécessitant plusieurs ressources. Ils sont en majorité para-médics et peuvent dispenser des soins.

Les ambulances et les véhicules-médecins sont les principales ressources. Les premières sont en plus grand nombre que les deuxièmes. Les activités de planification concernent particulièrement ces deux types de ressource. Essentiellement, les activités d'Urgences Santé sont divisées en deux champs : les services inter-établissements et les appels d'urgence de type 4242.

1.3.3 Services inter-établissements

Les services inter-établissements consistent à transporter des patients d'un établissement vers un autre (CH-CH, domicile du patient-CH, CH-domicile du patient). Un transport de ce type peut avoir lieu dans le cas où une personne invalide ne peut se rendre de son domicile au CH pour un examen, et inversement. Mais dans la plupart des cas, c'est pour des raisons d'équipements spécialisés qui ne se trouvent que dans certains établissements. Si des patients résident dans des établissements qui n'ont pas l'équipement nécessaire pour leur traitement, ils devront se déplacer vers des établissements plus spécialisés. Le retour au centre hospitalier d'origine génère également un transport inter-établissement.

Pour ce type de service, il est possible d'appeler à l'avance pour en faire la demande. La nature de ces appels est différente de celles des appels d'urgence ordinaires : certains sont connus à l'avance (25%) et une grande majorité de ceux-ci (95% ?) ne sont pas urgents. Il faut dire qu'à l'origine, le patient est, dans la plupart des cas, dans un centre hospitalier et bénéficie d'un entourage adéquat. Le service n'est actif que les jours de semaine, en particulier entre 7:30 et 17:00, c'est-à-dire durant les heures de travail des employés spécialisés des centres hospitaliers.

Le fait de connaître des demandes à l'avance permet de jumeler des appels en affectant des ambulances à plusieurs civières. Aussi, on peut "chaîner" des appels et ainsi faire un itinéraire de travail pour un même véhicule. Par exemple, soient deux transports devant être effectués successivement dans le temps : le premier du CH no 1 au CH no 2 et l'autre du CH no 2 au CH no 3. On peut alors affecter une seule ambulance qui fera les deux transports.

1.3.4 Services d'urgence (appels de type 4242)

Un appel de type 4242 est un appel d'urgence classique. On le nomme ainsi en raison du numéro de téléphone unique (842-4242) que toute personne sur le territoire peut composer pour obtenir les services d'une ambulance. Pour un appel de ce type, une ou plusieurs unités (ambulance et/ou véhicule-médecin) peuvent se rendre sur les lieux d'une urgence médicale pour transporter le ou les patient(s) à l'hôpital. Pour se faire une meilleure idée de la situation, voyons comment les différents événements s'enchaînent lors d'un tel appel.

Chronologie des événements

1) Une urgence médicale (accident avec blessé, arrêt cardiaque, ...) se produit sur le territoire de l'île de Montréal ou de l'île Jésus.

2) L'urgence est signalée par un appel téléphonique; tous les appels sont centralisés et il n'y a qu'un seul numéro à composer (842-4242).

3) Un membre du personnel répond à l'appel. Il possède des qualifications médicales (c'est un(e) infirmier(e) ou un médecin) et il est en mesure de filtrer les appels non pertinents. Il recueille toutes les informations nécessaires (adresse, no. de tél., nombre de patients, leur état) pour l'envoi de ressource(s) sur les lieux. Il décide quel(s) type(s) de véhicule sont envoyé(s).

4) Après la prise des informations, la demande est acheminée à un répartiteur. Celui-ci consulte une grande carte murale magnétique indiquant la localisation des véhicules disponibles. Compte tenu de la provenance de l'appel et de son niveau d'urgence, il décide alors s'il envoie des ressources, et si oui, lesquelles.

Le répartiteur ne bénéficie pas d'un système de localisation automatique des véhicules. Il connaît exactement l'emplacement des véhicules immobiles, mais seulement de façon approximative celui des véhicules en mouvement. Un tel système automatique permettrait d'afficher en tout temps, sur un tableau lumineux par exemple, la localisation exacte des véhicules (immobiles ou non) de même que leurs états respectifs (disponible ou non).

5) Une fois arrivés sur les lieux de l'urgence, les ambulanciers s'assurent de l'état des patients et évaluent la situation. Ils peuvent alors demander l'envoi de ressources supplémentaires si le besoin s'en fait sentir.

Ils donnent les premiers soins aux patients et les stabilisent, puis ils les transportent à l'hôpital si cela s'avère nécessaire. Une fois leur intervention terminée, ils doivent remplir des formalités administratives avant de redevenir disponibles.

Note : Durant toute la période entre l'affectation de l'unité et son retour en disponibilité, l'ambulancier est en contact radio avec un des répartiteurs. Il lui transmet, entre autres choses, l'heure exacte à laquelle survient chacune des étapes du processus d'intervention. C'est également avec lui qu'il tentera de préciser une adresse ou encore d'obtenir des informations sur la nature du cas à traiter.

1.3.5 Particularité de la situation à Montréal

Dans un système d'urgence, en général, les véhicules affectés à la couverture d'un territoire se trouvent en des endroits fixes et déterminés lorsqu'ils sont en disponibilité. Cela signifie qu'à la fin de tout appel, lors d'un retour en disponibilité, un véhicule donné retourne toujours à son point de disponibilité dans l'attente d'un prochain appel. La plupart du temps, ces points sont des lieux aménagés offrant divers services (ex : caserne offrant douche, cuisine, lits, divertissements, etc.).

On retrouve notamment de tels postes d'attente pour les camions de pompiers qui sont ordinairement localisés dans des casernes. C'est aussi le cas pour les ambulances dans certaines villes américaines et à Toronto. À Montréal cependant, il n'existe pas de telles casernes (ou garages) où les unités disponibles seraient localisées. Les ambulances et les véhicules-médecins peuvent donc, en principe, être localisés n'importe où sur le territoire en attendant de répondre à une urgence. De plus, un véhicule en disponibilité ne retourne pas nécessairement au même endroit où il était avant son affectation à un appel. Cela dépend de l'état du système au moment de son retour en disponibilité. Enfin, il est possible de relocaliser des unités n'importe où sur le territoire.

1.4 Motivations du projet du C.R.T.

Afin d'offrir un service de qualité il est nécessaire de déterminer un nombre suffisant de ressources à affecter. Toutefois, le coût annuel d'opération d'une ambulance vingt-quatre heures par jour est très élevé (supérieur à 250 000 $). Il importe donc de planifier adéquatement les ressources pour ne pas en affecter beaucoup plus qu'il n'en faut.

D'un autre côté, le territoire à couvrir est grand et les taux d'appel pour l'ensemble de celui-ci sont déjà très élevés et ne cessent de

croître (en juin 1984, le nombre moyen d'appels par jour était supérieur à 500). Ainsi, une planification manuelle des ressources et des opérations est déjà difficile. Si l'on vise une utilisation optimale des ressources, cette planification est encore plus difficile à accomplir.

Compte tenu de la complexité grandissante du problème et des difficultés qui en résultent, Urgences Santé a demandé au Gouvernement québécois des fonds pour l'implantation d'un système de communication moderne assisté par ordinateur. Un tel système permettrait d'obtenir un ensemble de données utiles à l'exploitation et à la planification. En plus de ce système, il serait grandement souhaitable que les gestionnaires puissent bénéficier de toute une gamme d'outils qui leur permettraient de prendre des décisions qui seraient basées sur ces données.

Parallèlement à cela, le Centre de recherche sur les transports (C.R.T.) a décidé, en 1984, de mettre de l'avant un projet dont le but était d'étudier les problèmes reliés à la gestion des ressources d'un système d'urgence ambulancier à Montréal. Ce projet vise à apporter des solutions utilisables dans la pratique. Il a été subventionné par le programme F.C.A.R. action spontanée du gouvernement du Québec. Le présent mémoire s'inscrit dans ce projet.

1.5 Buts du présent mémoire

Nos efforts se concentreront, dans ce mémoire, sur les problèmes reliés essentiellement au déploiement des ambulances affectées au service d'urgence (appels de type 4242). Le service inter-établissement ne sera pas considéré, pas plus que les types de ressource autres que les ambulances.

Les performances d'une flotte d'ambulances sont étroitement liées au temps de réponse. Tout gestionnaire est intéressé à connaître la valeur d'une stratégie donnée a priori, c'est-à-dire sans l'essayer en pratique.

Afin d'étudier différentes stratégies de déploiement, il est nécessaire de disposer d'un outil d'évaluation des performances du système (temps de réponse moyen global, temps de réponse maximal, temps de réponse par secteur, ...). Cela permet d'analyser différentes possibilités. Par exemple, on peut s'intéresser à la localisation des postes d'attente des ambulances lorsque celles-ci sont en disponibilité. Avant d'implanter une stratégie donnée, on voudra avoir une idée des performances pour plusieurs choix de localisation sans être obligé de les expérimenter sur le terrain.

Un de nos principaux buts consistera donc à étudier quelques outils d'évaluation d'un système d'urgence. Étant donné la nature stochastique d'un tel système, nous nous pencherons en particulier sur un modèle de file d'attente (modèle hypercube) ainsi que sur des résultats de simulation. Le modèle de file d'attente utilise un certain nombre d'hypothèses. Nous nous intéresserons à voir si les résultats qu'il fournit sont encore valables lorsqu'on s'éloigne des hypothèses pour se rapprocher de ce qui se passe dans la réalité. La simulation servira à mettre le modèle à l'épreuve.

Nous aborderons, au chapitre 2, la description des mesures des performances, de même que celle des objectifs que l'on vise dans le déploiement d'une flotte d'ambulances. Au chapitre 3, nous esquisserons une revue de certaines connaissances actuelles dans le domaine des systèmes d'urgence. Les chapitres 4 et 6 seront quant à eux consacrés à l'étude du modèle hypercube. On s'intéressera en particulier à comparer les performances calculées à partir de ce modèle à celles obtenues à l'aide d'un programme de simulation pour différentes situations s'éloignant des hypo-

thèses du modèle. Au chapitre 5, une analyse statistique sera faite pour déterminer les différents paramètres utilisés dans le modèle analytique et dans la simulation. Cette analyse statistique permettra, en outre, de donner une description de la réalité.

CHAPITRE II

MESURES DES PERFORMANCES ET OBJECTIFS

2.1 Généralités

La description des problèmes reliés au déploiement d'une flotte d'ambulances nécessite l'introduction d'un certain nombre de concepts. Entre autres, nous avons discuté d'une manière informelle de ce qu'est un système d'urgence. Avant d'aller plus loin, nous allons bien définir ce terme pour savoir exactement de quoi il est question.

La notion de système d'urgence intervient dans des situations qui sont essentiellement caractérisées par les trois points suivants :

- Des incidents se produisent un peu partout sur le territoire; le moment et le lieu exact de chacun de ces incidents ne sont pas connus à l'avance. Ces incidents donnent lieu à des appels d'urgence.

- En guise de réponse :
  - . si l'appel est prioritaire, des unités doivent immédiatement être affectées sur les lieux de l'urgence;
  - . si l'appel n'est pas prioritaire, on peut attendre avant d'affecter des unités.

- La rapidité avec laquelle les unités arrivent sur place, en particulier pour les appels prioritaires, a un lien certain avec la qualité du service à offrir.

Par conséquent, un tel système se définit comme l'ensemble des moyens mis en oeuvre pour réagir à ces situations.

Pour discuter de la problématique d'un système d'urgence, on doit connaître les objectifs que l'on vise, de même que les mesures utilisées pour les formuler. C'est ce à quoi nous allons maintenant nous attarder.

2.2 Description et justification des mesures utilisées

2.2.1 Utilité des mesures

Les principaux buts que se fixent les autorités consistent, d'une part à offrir un bon service à tous les usagers d'une manière équitable, et d'autre part, à planifier et à exploiter le plus efficacement possible l'ensemble des ressources à affecter à la couverture. Dans l'optique de l'utilisation des méthodes quantitatives pour atteindre ces buts, on se doit de définir certaines mesures qui serviront à quantifier les objectifs. Cependant, ces mesures ne tiendront compte que d'un sous-ensemble des particularités du système. En effet, comme on s'en doute, il sera impossible, par exemple, de mesurer combien de vies seront sauvées avec l'application de telle ou telle politique.

2.2.2 Temps de réponse

Cette mesure permet de donner une appréciation de la qualité du service. En effet, un temps de réponse plus court permet une intervention plus rapide auprès des blessés et des malades à transporter. Il importe dans les cas critiques d'intervenir le plus rapidement possible. Un temps de réponse trop long peut avoir des conséquences néfastes (décès, blessures traitées tardivement et entraînant des séquelles permanentes).

Idéalement, le temps de réponse se définit comme la différence entre le moment où l'urgence se produit (accident avec blessés, malaise soudain, ...) et l'heure à laquelle les services arrivent sur les lieux. Évidemment, en pratique, des données pour un tel temps de réponse sont impossibles à obtenir. Celles dont nous allons disposer seront nécessairement biaisées.

Nous allons d'abord chercher à définir un temps de réponse usuel sur lequel nous pourrons nous baser pour avoir une juste appréciation de la vraie mesure. Le tout est précisé à la figure 1.

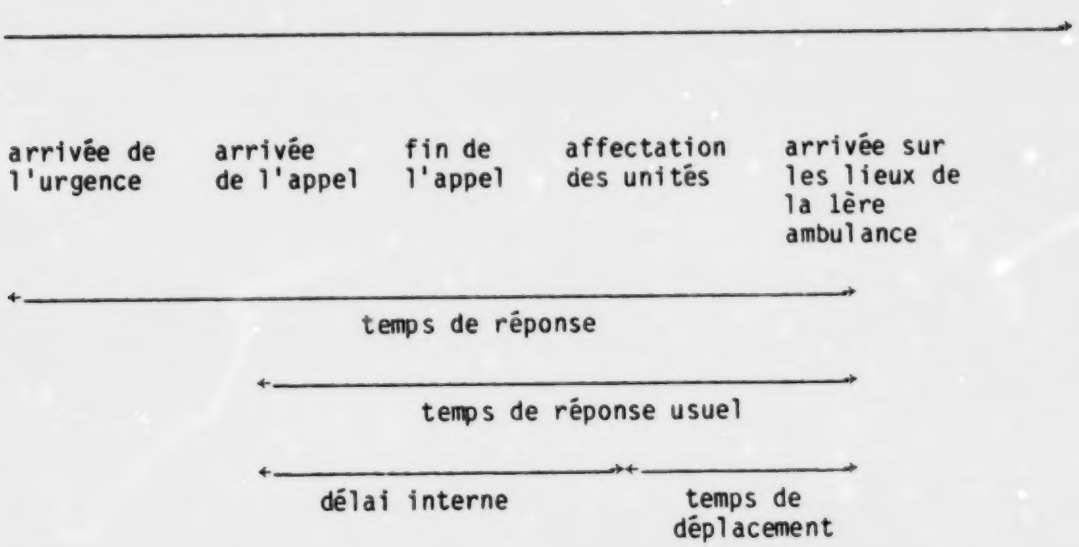


FIGURE 1

On distingue le moment de l'arrivée de l'urgence et celui de l'arrivée de l'appel. Le premier correspond au moment où l'urgence se produit sur le territoire alors que le deuxième est l'heure où l'appel est reçu chez Urgences Santé. Pour nous, le temps de réponse usuel sera le délai entre l'heure de l'arrivée de l'appel et l'heure de l'arrivée sur les lieux de la première ambulance. Ce temps est la somme du délai interne (temps mis pour recueillir les informations et pour trouver les unités disponibles à envoyer) et du temps de déplacement de l'ambulance.

2.2.3 Charges de travail

Dans la recherche de méthodes de planification et d'exploitation d'une flotte de véhicules ambulanciers, il est utile de savoir si les ressources que l'on affecte sont suffisantes. Par ailleurs, on s'intéresse également à ce que les unités soient suffisamment utilisées (et pas trop). Dans ce cas, on aura besoin de connaître le taux d'utilisation de chacune d'elles. C'est ce qu'on appelle la charge de travail (ρ) et qui se définit comme suit :

> charge de travail de l'unité i = $\rho_i \triangleq$ proportion du temps où l'unité i est occupée à servir un appel par rapport au temps total où elle est en service.

Une des préoccupations peut alors consister à éviter les situations où certains véhicules sont toujours occupés pendant que d'autres ne le sont presque pas. On cherchera alors à équilibrer les charges de travail.

2.3 Objectifs

Comme nous l'avons déjà dit, un des buts consiste à offrir le meilleur service à tous les usagers, et ce, d'une manière équitable. Il est nécesssaire alors de définir des critères de qualité de service.

Une urgence médicale nécessite une rapidité d'intervention de la part des secours. D'une manière générale, on voudra minimiser le temps de réponse. En pratique, pour une période donnée, on visera à minimiser le temps moyen de réponse.

Toutefois, l'application de mesures visant uniquement à minimiser le temps moyen de réponse peut inclure des temps de réponse très élevés pour des régions à faible taux d'appel. Cela pose un problème, car on

voudrait bien que toute personne, où qu'elle soit sur le territoire, puisse bénéficier en tout temps d'une intervention rapide en cas d'urgence. C'est là qu'intervient la notion d'équité. On peut l'exprimer en fonction du temps maximal de réponse. Par exemple, aux États-Unis, le standard d'équité consiste à ce que 90% des appels aient un temps de réponse inférieur ou égal à dix minutes.

Le critère de qualité de service consiste donc à minimiser le temps moyen de réponse sujet à une contrainte de borne supérieure sur le temps de réponse. Plus formellement, soit T le temps de réponse, alors l'objectif consiste à

$$\text{(P)} \qquad \begin{aligned} &\text{Minimiser } E[T] \\ &\text{s.a.} \quad P[T < \beta] \geq 1-\alpha \qquad 0 \leq \alpha \leq 1 \\ &\qquad\qquad\qquad\qquad\qquad\quad \beta > 0 \end{aligned}$$

(par exemple α=0.1 et β=10 minutes)

En pratique, on voudra déterminer des politiques de déploiement pour un nombre donné de ressources de façon à satisfaire au maximum les critères de qualité de service. Cependant, on s'intéresse également à déterminer le nombre minimal de ressources à affecter, tout en gardant satisfaits les même critères de qualité de service (car on vise une utilisation rationnelle des ressources).

Il est à remarquer que le problème (P) ne permet pas de résoudre tous les problèmes d'équité. C'est ainsi que nous n'avons aucun contrôle sur le degré de violation de la fraction (au plus α) des appels ne respectant pas la contrainte de temps maximal. On peut donc raffiner (P) en ajoutant des contraintes restreignant de plus en plus les écarts permis. Par exemple, avec β=10 et α=0.10, on pourrait ajouter une deuxième contrainte où β'=15 et α'=0.05 pour resserrer l'étau.

On peut également augmenter le raffinement en tenant compte de différents types d'appel (par exemple, ceux de différentes priorités) en spécifiant des objectifs indépendants pour chaque type.

2.4 Aspects aléatoires du système

Dans un système d'urgence, on rencontre plusieurs aspects aléatoires. Entre autres, il y a évidemment l'arrivée de la demande dans le temps et dans l'espace, de même que le temps de service(1). Si on ajoute à ces aspects le fait que des unités occupées ne peuvent répondre à d'autres appels, il en résulte une incertitude quant à la disponibilité de chacun des véhicules. C'est ce que l'on appelle l'aspect aléatoire de l'état du système et il est très important. On devra donc s'efforcer de tenir compte le plus possible des phénomènes aléatoires dans les techniques de résolution.

Ainsi, on devra faire attention de ne pas simplifier la situation en utilisant des méthodes déterministes dans les cas où l'activité du système est élevée. Par exemple, aborder le problème de la localisation seulement en fonction des demandes et des distances (approche déterministe), et ce, sans considération de l'aspect aléatoire de l'état des véhicules (disponible ou non), est une approche qui n'est appropriée que si le taux d'appel est très faible (garantissant ainsi la disponibilité de toutes les ressources presqu'en tout temps).

Dans le cas où la fréquence des appels est élevée, on devra pouvoir prendre en considération la détérioration du service; cette détérioration étant issue de la non-disponibilité des unités pour répondre aux appels et du phénomène de congestion qui en résulte.

---

(1) Le temps de service d'un véhicule affecté à un appel se définit comme la différence entre le moment où il est affecté et le moment où il redevient disponible une fois qu'il quitte le patient (et non une fois revenu en son point de localisation).

Il est donc nécessaire, pour nous, d'avoir des outils permettant d'évaluer les performances du système sous une configuration particulière, étant donnée la nature dynamique de l'état des véhicules. C'est l'objet des modèles de files d'attente et des techniques de simulation. Dans le prochain chapitre (3), nous allons passer en revue quelques-unes des méthodes développées jusqu'à ce jour pour évaluer les performances des systèmes d'urgence et pour apporter des solutions aux différents problèmes du déploiement.

CHAPITRE III

SUR L'ÉTAT DES CONNAISSANCES ACTUELLES DANS LE DOMAINE DES SYSTÈMES D'URGENCE

3.1 Généralités

Nous avons vu, au chapitre précédent, l'importance de la prise en considération de l'aspect stochastique dans l'analyse des opérations d'une flotte de véhicules d'urgence. Il existe plusieurs ouvrages sur les processus stochastiques. Notre but n'est pas d'en faire la revue. Cependant, nous pouvons référer le lecteur à Feller [1] et à Heyman-Sobel [2]. Parmi les modèles stochastiques, ceux ayant trait aux files d'attente vont particulièrement retenir notre attention. À ce sujet, Kleinrock [3,4] constitue une bonne référence.

Dans le domaine particulier de l'application des modèles stochastiques de files d'attente aux systèmes d'urgence, Larson [5,6] a écrit deux ouvrages de référence. Entre autres, "Urban Operations Research" écrit en collaboration avec Odoni [6] est celui le mieux adapté pour se familiariser avec notre sujet.

Pour pouvoir analyser la nature d'un système d'urgence et développer des stratégies de déploiement, des outils mathématiques et informatiques sont évidemment indispensables. Nous diviserons ces derniers en deux classes : ceux servant à évaluer les performances du système et ceux qui fournissent des solutions à des problèmes particuliers du déploiement. Dans ce qui suit nous allons aborder brièvement la revue de ceux qui nous semblent les plus importants.

3.2 Outils d'évaluation des performances du système

Le but de ces outils est d'évaluer les performances d'un système d'urgence (temps moyen, temps maximal, charges de travail, etc.) en fonction

1) de la répartition de la demande dans le temps et dans l'espace,
2) du nombre d'unités et de leur localisation,
et 3) d'une politique de répartition des véhicules pour les assigner aux demandes particulières des usagers.

On rencontre ici deux types de méthode : les techniques de simulation et les modèles analytiques. Dans notre approche, nous allons utiliser les deux (nous verrons pourquoi au chapitre 4).

Toutefois, nous ne discuterons pas en détail de la conception d'un modèle de simulation adapté à la situation d'Urgences Santé. Celui que nous allons utiliser a été développé au Centre de recherche sur les transports. Le lecteur intéressé à la simulation en général peut néanmoins consulter Law et Kelton [7] de même que Bratley, Fox et Schrage [8].

Depuis une quinzaine d'années, le domaine des modèles analytiques appliqués aux systèmes d'urgence s'est développé. En particulier, une étude sur le déploiement des véhicules de pompier à New-York a donné lieu à plusieurs articles. Mentionnons [9,10,11,12,13,29] pour n'en citer que quelques-uns. Parmi les résultats produits dans cette étude, un modèle simple de prédiction retient notre attention [12]. Ce dernier prédit le temps moyen de déplacement dans une région en fonction de l'aire du territoire, du nombre de véhicules affectés à la couverture, du taux d'appel et du temps de service moyen.

On se tournera plutôt vers la classe des modèles de files d'attente. Il est à remarquer que dans le déploiement d'une flotte de véhicules d'urgence, on doit tenir compte d'une situation qui évolue dans le temps

et dans l'espace. Les modèles classiques de file d'attente décrivent des processus se déroulant uniquement dans le temps. Il existe toutefois un modèle de files d'attente où la répartition des demandes dans l'espace est prise en considération : c'est le modèle de files d'attente hypercube décrit par Larson [6,14,15,16]. Dans ce modèle, l'aspect stochastique de la disponibilité des véhicules d'urgence est pris en ligne de compte. Il s'ensuit, évidemment, une description du phénomène d'attente en fonction de l'évolution du système dans le temps et dans l'espace.

Nous ferons plus loin une description plus détaillée de l'hypercube, car il constitue une composante essentielle de notre étude. Nous nous proposons d'analyser les résultats qu'il fournit et d'examiner dans quel sens il pourrait être modifié pour mieux correspondre au problème tel qu'il se pose à Montréal. Nous pouvons tout de suite attirer l'attention du lecteur sur les difficultés d'une telle adaptation. Ainsi, dans l'hypercube, on ne tient pas compte de la priorité des appels, ni de la relocalisation des véhicules d'urgence. De plus, pour chaque appel, on suppose qu'un seul véhicule est envoyé sur les lieux. Ce qui nous importe, c'est de voir si les résultats qu'il fournit se rapprochent des performances réelles de façon à ce qu'on puisse l'utiliser. On répondra en partie à cette question.

Les modèles dont nous venons de discuter servent à décrire la nature du système, de même qu'à estimer les différentes performances qui lui sont caractéristiques. En aucun cas, il ne faut voir ces modèles comme fournissant des solutions optimales à des problèmes particuliers du déploiement. Ils ne servent qu'à calculer les performances du système pour des politiques de déploiement qui doivent être spécifiées a priori. Dans la prochaine section, nous verrons les outils servant à obtenir des solutions.

3.3 Modèles de recherche de solutions

On rencontre plusieurs types de modèle propres aux divers aspects du problème du déploiement. Nous traiterons de ceux concernant la localisation, la répartition et la relocalisation.

3.3.1 Modèles de localisation

Plaçons nous dans le contexte où des demandes pour des biens et des services se produisent d'une façon aléatoire sur un territoire. D'une manière générale, le but de ces modèles de localisation est de déterminer, d'une part, le nombre de ressources nécessaires pour ce territoire, et d'autre part, les emplacements où devront être localisées ces ressources. Pour atteindre ce but, ils suggèrent des localisations optimales pour un nombre spécifié de ressources à affecter sur le territoire afin de satisfaire des critères de qualité de service donnés.

3.3.1.1 Cas où les ressources sont toujours disponibles (contexte déterministe)

Nous qualifierons de déterministes les problèmes de localisation où l'état de chacune des ressources (disponible ou non) n'est pas sujette à une loi de probabilité. Dans de tels cas, les ressources sont supposées comme étant toujours disponibles. Une telle situation n'est évidemment pas la nôtre.

Note : Même si la demande reste un phénomène aléatoire, nous tenons quand même à qualifier cette classe de problème comme étant déterministe à cause des méthodes utilisées pour les résoudre. Celles-ci ne font pas appel à des notions de processus stochastiques.

Illustrons maintenant ce contexte à l'aide d'un exemple. Supposons que l'on veuille trouver les emplacements optimaux de n entrepôts dans un contexte donné. Supposons aussi que ces derniers possèdent en quantité toujours suffisante les biens qu'ils ont à distribuer. On remarque que

l'on peut alors les considérer comme des ressources continuellement disponibles. Le problème consiste à trouver des emplacements pour ces n entrepôts afin de satisfaire les clients en un temps raisonnable. Nous allons nous intéresser brièvement à la classe des problèmes de ce type.

En supposant que l'on utilise une structure de réseau pour décrire le territoire, il existe alors des méthodes de solutions analytiques pour la classe des problèmes ci-haut mentionnée. Handler et Mirchandani [17] en ont fait une excellente revue. Pour l'essentiel, disons que deux types principaux de modèles se présentent : le "n-médiane" et le "n-centre". Les deux sont reliés à des critères d'optimalité différents.

Le problème "n-médiane" consiste à déterminer les emplacements optimaux de n ressources, que l'on suppose toujours disponibles, dans le but de minimiser le temps moyen de réponse. Le problème "n-centre" vise plutôt à déterminer les emplacements optimaux afin de minimiser le temps maximal de réponse. Différentes approches peuvent être utilisées pour résoudre ces problèmes. L'exposé de celles-ci est fait dans [17].

On peut aussi trouver des solutions dans le cas où l'on aborde ensemble les deux objectifs précédents. On peut alors chercher à minimiser une combinaison convexe des temps moyen et maximal [31], ou encore, minimiser le temps moyen de réponse sujet à une contrainte de borne supérieure sur le temps maximal. On remarquera que cette dernière formulation correspond à celle des critères de qualité de service cités au chapitre 2. Une formulation mathématique en problème de programmation en nombres entiers en a été donnée par Toregas, Swain, Revelle et Bergman [18].

Il est à noter, encore une fois, que ces méthodes de solution se situent dans un contexte déterministe dans la mesure où chacune des ressources est toujours disponible. Cela permet d'utiliser une règle de répartition simple qui consiste à toujours envoyer la ressource la plus

près. Dans le cas où les ressources ne sont plus disponibles en tout temps, de telles solutions ne tiennent plus. On doit alors tenir compte, en particulier, des congestions dues au phénomène d'attente. On peut cependant utiliser les techniques de localisation en contexte déterministe pour générer de bonnes solutions initiales ou encore dans une situation où les taux d'appel sont faibles (dans ce cas chaque ressource est "presque" toujours disponible).

3.3.1.2 Cas où les ressources ne sont pas toujours disponibles (contexte non déterministe)

Dans un système d'urgence ambulancier, il se peut qu'une unité soit occupée lorsque se produit un appel dans le voisinage de son point de localisation. On doit alors nécessairement prendre en considération le phénomène "d'engorgement" des appels produit par la non-disponibilité des unités de service. Cela a été étudié en particulier pour étendre le problème n-médiane au contexte non-déterministe (on l'appelle n-médiane avec congestion).

Berman et Larson [19] ont en effet étudié ce problème. Sous certaines hypothèses, ils ont démontré l'existence d'une solution sur les sommets du réseau. De plus, Jarvis [20] a développé une procédure qui trouve des solutions localement optimales. Dans cette procédure, il utilise le modèle hypercube.

3.3.2 Relocalisation

La relocalisation constitue un problème complexe. Dans une modélisation de cette situation, on doit tenir compte que dans une relocalisation, où plusieurs véhicules sont en mouvement, il peut survenir un ou plusieurs appels. Les unités en mouvements restent disponibles et on doit considérer cet aspect. Les modèles développés jusqu'à maintenant utilisent des hypothèses simplificatrices.

Il est possible d'obtenir une stratégie de relocalisation dans un contexte idéalisé qui pourra constituer une borne inférieure sur la solution optimale de la fonction objective que l'on veut minimiser. Jarvis [20] discute de cette idée. Celle-ci est simple. Il suffit de supposer que la relocalisation peut être effectuée instantanément. Dans un tel cas, il est évidemment toujours payant de relocaliser à chaque fois que la situation se présente.

Dans [13], on discute d'une méthode en temps réel développée pour relocaliser les camions de pompiers dans la ville de New York. On utilise, ici, des arguments de programmation mathématique ("set covering") sans toutefois tenir compte du phénomène d'attente d'une manière explicite.

Plus récemment, Berman [21] a innové en développant une méthode analytique basée sur la théorie des processus markoviens. Il considère q véhicules identiques initialement localisés en q sommets d'un réseau. Ici, l'état du système n'est pas défini en termes des états propres à chacun des q véhicules (disponible ou non), mais plutôt en termes des états des q localisations (occupée ou non). En ce sens, cette approche est donc différente de celle utilisée dans l'hypercube. Essentiellement, Berman traite le problème de la relocalisation en développant une façon de calculer le coût associé au déplacement d'une seule unité à la fois, et ce, d'un point i vers un point j sur le réseau. Les hypothèses sont cependant assez fortes (arrivées de Poisson pour les demandes, temps de service issus d'une loi exponentielle, temps de déplacement minime par rapport au temps passé sur les lieux, pas plus d'un événement (appel ou fin de service) peut arriver durant une relocalisation, retour en disponibilité vers l'emplacement le plus près). Berman identifie, dans ce contexte, la meilleure relocalisation (d'un seul véhicule) à faire, et ce, pour tous les états possibles du système.

3.3.3 Répartition

Dans une situation déterministe, on envoie toujours les ressources les plus près, car celles-ci sont disponibles en tout temps. Or, dans un contexte stochastique ce n'est plus le cas et on doit spécifier les unités qui seront affectées. Pour nous, une politique de répartition consiste essentiellement à déterminer quel véhicule, parmi ceux disponibles, sera affecté à un appel donné. On rencontre ici la notion de politique avec préférences fixées ("fixed preference dispatching"). Une telle politique consiste à désigner à l'avance, pour chaque zone du territoire, une liste de véhicules ordonnée selon la préférence. Étant donnée une zone particulière, le premier élément de la liste sera le véhicule préféré parmi tous les autres pour répondre à un appel dans cette zone. Or, ce véhicule préféré n'est pas nécessairement disponible en tout temps. Ainsi, lorsque surviendra un appel, on enverra le premier véhicule disponible que l'on rencontrera en balayant la liste.

Intuitivement, on peut penser que la meilleure politique de type "préférences fixées" est celle qui consiste à toujours désigner l'unité la plus près parmi celles disponibles. Carter, Chaiken et Ignall [22] ont rigoureusement démontré, pour le cas de deux véhicules, que ce n'est pas la politique du plus près qui est toujours la meilleure. Comme autre résultat, ils ont trouvé une procédure qui détermine, toujours pour deux véhicules, la frontière optimale entre deux zones d'affectation de telle sorte à minimiser la distance moyenne parcourue ainsi qu'à équilibrer les charges de travail.

Finalement, Jarvis a étudié le problème de la répartition pour un nombre N de véhicules [20, chap. 3]. Il ressort de son étude que, pour la plupart du temps, la politique optimale de répartition n'est pas celle du plus près. En général, cette politique optimale n'est même pas de type

"préférences fixées" (c'est-à-dire qu'elle varie avec l'état du système). Toutefois, la politique du "plus près" se situe à l'intérieur d'un pourcent de la politique optimale (selon des résultats empiriques de Jarvis).

CHAPITRE IV

DESCRIPTION DU MODÈLE HYPERCUBE ET DE LA MÉTHODOLOGIE

4.1 Buts

Une des préoccupations premières dans l'étude d'un système d'urgence consiste à se donner des outils d'évaluation des performances (temps de réponse moyen, charges de travail, etc.). Ici, la simulation constitue un premier recours. C'est en effet une des méthodes les plus souvent utilisées en recherche opérationnelle pour des situations non-déterministes et complexes à modéliser (comme un système d'urgence).

Dans un programme de simulation, l'ordinateur est utilisé afin d'imiter les opérations du système. Des appels sont générés dans le temps et dans l'espace de la même façon qu'ils ont été observés dans la réalité. En réponse à ces appels, on simule des temps de déplacement et des temps de service pour les unités affectées. Celles-ci sont déployées selon des politiques déterminées. Ce procédé permet de recueillir, pour chaque appel simulé, les informations nécessaires au calcul des différentes statistiques d'intérêt.

La simulation est un outil assez souple en ce sens qu'il est possible d'obtenir des performances pour à peu près n'importe quelles situations qui vont des plus simples (aucune priorité, aucune relocalisation) aux plus compliquées (plusieurs priorités, plusieurs relocalisations). Toutefois, un programme de simulation utilisé dans une situation complexe, comme celle d'Urgences Santé, nécessitera beaucoup de temps machine et d'espace mémoire. S'il est utilisé sur une base régulière, comme outil d'exploitation, son coût devient assez élevé.

De là découle l'intérêt de développer des outils analytiques. Ils permettent, à partir d'une modélisation mathématique du système, d'obtenir analytiquement des formules pour le calcul de certaines performances. Les statistiques ainsi obtenues nécessiteront peu de ressources ordinateur comparativement à la simulation. Cependant, on devra payer ailleurs pour cette économie réalisée. En effet, étant donné qu'un système d'urgence est très complexe à modéliser, on devra poser des hypothèses très simplificatrices. C'est le cas notamment du modèle hypercube.

Ultimement, nous aimerions vérifier si le modèle hypercube pourrait être adapté et utilisé dans le contexte d'Urgences Santé. Nous serons cependant plus modestes. Nous mettrons à l'épreuve le modèle de Larson en nous éloignant graduellement des hypothèses qu'il nécessite sans se rendre toutefois jusqu'au contexte réel.

La méthode consistera à comparer les statistiques obtenues analytiquement à partir du modèle avec celles calculées à l'aide d'un programme de simulation. Ce dernier reproduira des situations s'éloignant graduellement des hypothèses de Larson.

Les situations simulées seront des simplifications du contexte d'Urgences Santé. Nous n'allons pas, dans le cadre de notre étude, nous rendre jusqu'à la réalité. On notera que dans une étude complète et de plus longue haleine, on se doit de vérifier si le modèle peut être viable dans la pratique. Cela implique entre autres que l'on doive examiner si l'état d'équilibre, que l'on suppose dans le modèle, peut être atteint dans un contexte réel. Un tel travail dépasse le cadre de ce mémoire et ne sera donc pas fait ici.

Enfin, mentionnons que les simulations que nous ferons ne seront pas nécessairement utiles en pratique. Il faut retenir que notre but, à ce niveau-ci, consiste à mettre le modèle hypercube à l'épreuve et non à produire des résultats immédiatement utilisables.

4.2 Description du modèle hypercube

4.2.1 Idée générale

Avant d'aller plus loin, il importe de donner une description succincte du modèle hypercube pour les besoins de l'étude. Le lecteur intéressé par un développement détaillé pourra se référer à [6,14,15,20].

L'hypercube est un modèle de file d'attente où les serveurs sont différenciés. Cette distinction découle de la prise en considération du fait que les serveurs (ou unités d'urgence) sont localisés en des endroits différents sur le territoire lorsqu'ils sont en disponibilité. L'hypercube est utilisé pour obtenir des statistiques sur les performances d'un système dans un contexte bien particulier défini par toute une gamme d'hypothèses que nous allons maintenant examiner.

4.2.2 Hypothèses du modèle

Il y a dix hypothèses principales. Nous allons les formuler en les situant dans notre contexte. Plus précisément, nous verrons lesquelles pourront être vérifiées.

1) Le territoire doit être partitionné en N secteurs. Pour chacun d'eux, on associe un centroïde d'où émanera la demande pour l'ensemble du secteur. Nous utiliserons une partition de 468 secteurs ayant été faite à partir des secteurs de recensement tels que définis par Statistiques Canada.

2) Les arrivées des appels se produisent dans chaque secteur $j,(j=1,\ldots,N)$, selon un processus de Poisson et à un taux λ_j par unité de temps. On devra, ici, vérifier cette hypothèse statistique.

3) Les distances entre les secteurs doivent être connues. Nous disposerons d'une table de distances entre les secteurs de recensement. Cette table a été construite à partir des informations contenues dans les "Areas Master Files" fournies par Statistiques Canada [28].

4) On considère n unités d'urgence affectées à la couverture du territoire. Chacune de ces unités peut se rendre au besoin n'importe où sur le territoire (c'est-à-dire en n'importe quel secteur).

5) Lorsqu'une unité est disponible, elle est localisée en un poste d'attente fixe qui correspond au centroïde d'un secteur. Chacune des unités est assignée à un poste d'attente unique (contexte garage). La situation à Urgences Santé est bien différente, car les unités peuvent en principe, être localisées n'importe où (il n'y a pas de garage). En plus, une même ambulance n'occupe pas toujours le même emplacement lorsqu'elle est en disponibilité.

6) En réponse à un appel, exactement une unité est envoyée lorsqu'au moins une est disponible. Si aucune n'est disponible, l'appel est plaçé dans une file d'attente de type FIFO ("first in, first out"). Lorsqu'une unité redevient disponible, elle retourne instantanément à son poste s'il n'y a pas d'appel en attente. Dans le cas contraire, elle est envoyée directement sur les lieux du prochain appel, sans passer par son garage.

On remarquera qu'en réalité, plus d'une seule ambulance peuvent répondre à un même appel.

7) La répartition des unités se fait selon une politique de type "fixed preference". Pour chacun des secteurs, on détermine à l'avance la liste des unités à affecter selon l'ordre de préférence. C'est à partir de la table des distances que l'on construira ces listes de préférence en prenant toujours, pour chaque secteur, l'unité la plus près d'abord. On procède d'une manière analogue à Urgences Santé en envoyant toujours l'ambulance disponible la plus près. Toutefois, cette politique n'est pas prédéfinie en termes de serveurs étant donné que ceux-ci, en pratique, n'occupent pas toujours les mêmes postes d'attente.

8) On suppose que le temps de service (affectation-retour en disponibilité) est réparti, pour chaque unité i, selon une loi exponentielle avec moyenne $1/\mu_i$. Cette hypothèse devra également être vérifiée.

9) La variabilité du temps de service s'explique en majeure partie par la variabilité du temps passé à répondre à l'appel (temps sur les lieux et temps à l'hôpital). L'apport du temps de déplacement (affectation-arrivée sur les lieux) est donc considéré comme minime. Toutefois ce n'est pas ce qui se passe en pratique. Les unités d'urgence mettent parfois du temps à rejoindre les patients.

Face à ce problème, il existe une procédure de calibration des temps moyens de service [6]. Celle-ci attribue à chaque unité i, un temps moyen de service $1/\mu_i$ qui tient compte du déplacement de chaque serveur, de même que du temps passé auprès du patient (sur les lieux et à l'hôpital).

10) Le système doit avoir atteint l'état d'équilibre ("Steady State"). Un tel état est atteint lorsque le système ne dépend plus des conditions initiales et que les probabilités des différents états ne varient plus avec le temps. En pratique, il n'est aucunement garanti que l'état d'équilibre est atteint en raison des taux d'appels qui fluctuent aléatoirement dans le temps.

Dans la réalité, il est évidemment impossible de rencontrer un système qui satisfasse toutes ces exigences. Il faut ajouter à cela que le modèle ne tient pas compte

i) du fait que l'on puisse traiter chaque appel selon une politique de priorité,

ii) que chacun des appels ne monopolise pas qu'une seule unité (pas d'affectation jumelée, triple ...),

iii) qu'il puisse y avoir des relocalisations

et iv) que les taux d'appel varient dans le temps.

Étant donné les hypothèses précitées, le modèle permet de calculer une grande quantité de statistiques. Compte tenu des objectifs (dont nous avons discuté au chapitre 2) les statistiques qui nous intéressent sont les suivantes :

Statistiques globales :
- temps moyen de déplacement (rapidité)
- P [temps de déplacement ≤ borne supérieure] (équité)

Statistiques par unité :
- temps moyen de déplacement
- charge moyenne de travail.

4.2.3 Formulation mathématique

4.2.3.1 Espace des états

Le modèle de file d'attente hypercube est en fait une désagrégation du modèle de file d'attente classique M/M/n (n serveurs, arrivée Poisson, temps service issus d'une loi exponentielle et de moyenne $1/\mu$ pour chaque serveur(1)).

Dans le modèle M/M/n, l'état des serveurs est défini par le nombre de serveurs occupés (l'état n est l'agrégation des états où il y a n serveurs d'occupés et k=0,1,2,... appels en attente). À l'état d'équilibre, on peut représenter ce modèle par ses états et par les taux de transitions entre chacun d'eux. La solution des équations à l'état d'équilibre est bien connue dans la documentation. Celle-ci donne les probabilités associées à chacun des états du système.

---

(1) Nous supposerons que le temps de service moyen est le même pour chaque serveur afin de faciliter la lecture du texte. En pratique, on considère des temps moyens de service différents à cause de la variabilité du temps de déplacement et surtout de l'importance relative de ce dernier par rapport au temps de service.

Dans le modèle hypercube, l'espace des états est plus compliqué, car les serveurs sont distincts. On devra tenir compte de cette distinction pour énumérer les états. Plus précisément, on définit d'abord l'état d'un serveur i par la variable b_i.

$$b_i = \begin{cases} 1 \text{ si le serveur i est occupé} \\ 0 \text{ sinon} \end{cases}$$

L'état du système est ensuite défini par la prise en compte de l'état de chacun des serveurs; on le note par le n-tuple $(b_1, b_2, b_3, \ldots b_n)$. On voit qu'il y a 2^n états possibles dans l'hypercube contre n seulement dans le modèle M/M/n.

(Note : Encore ici, on agrège les états où il y a une file d'attente, c'est l'état (1,1, ..., 1)).

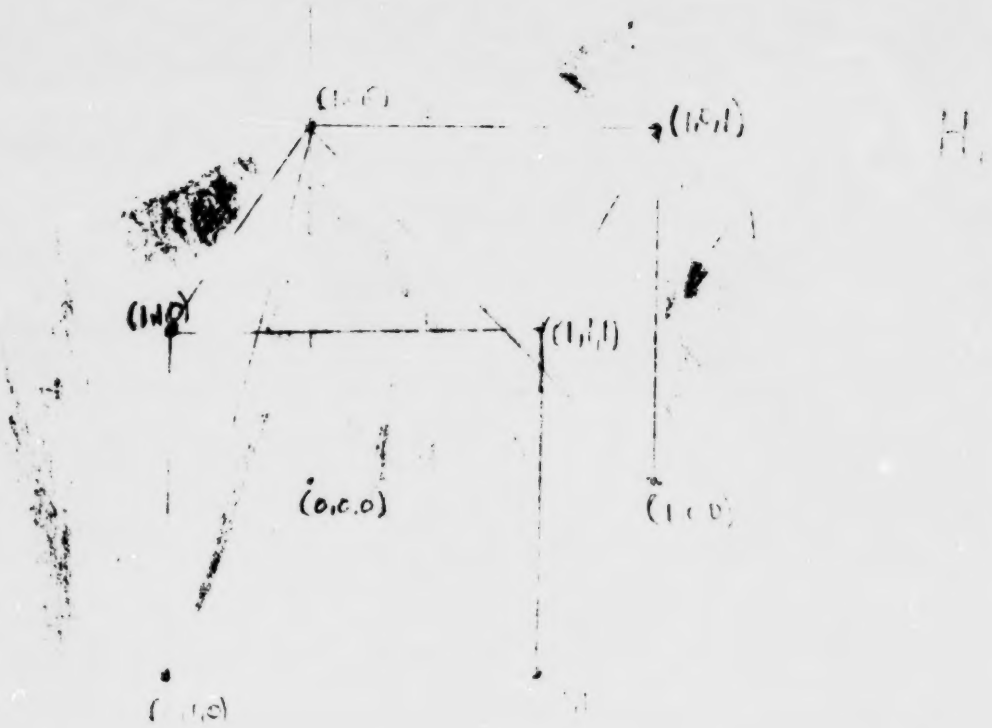
Soit H_k = l'ensemble des états pour lesquels k unités sont occupés $k = 1, \ldots, n$.

$$= (b_1, b_2, \ldots, b_n); \sum_{i=1}^{n_1} b_i = k \quad , k=1,\ldots,n.$$

On voit que H_k correspond en fait à un état du modèle M/M/n. H_k est un regroupement d'états de l'hypercube. En ce sens, l'espace des états dans l'hypercube est une désagrégation de celui de M/M/n.

On peut illustrer cela en considérant un exemple à trois serveurs. L'espace des états de l'hypercube est (0,0,0), (1,0,0), (1,0,1), (1,1,0), (0,0,1), (0,1,0), (0,1,1), (1,1,1).

En examinant la figure 4.2 qui montre l'espace des états, on remarque que H_k (k=1,2) est un hyperplan qui intersecte un hypercube dont les sommets sont les états du modèle désagrégé. C'est de là que vient le nom d'hypercube.



Espace des états de l'hypercube et hyperplan H_1

FIGURE 4.2

4.2.3.2 Taux de transition entre les états

On s'intéresse aux taux de transition entre les états. C'est à partir de ceux-ci que l'on peut obtenir les équations à l'état d'équilibre dont la solution permet de connaître les probabilités associées aux états du système.

Une des hypothèses stipule qu'une seule unité est affectée à la fois pour répondre à un appel. De plus, la probabilité que deux unités redeviennent libres au même moment est supposée pratiquement nulle. En conséquence, les transitions ne s'opèreront qu'entre états adjacents sur l'hypercube (voir Larson [6] p. 309). En d'autres mots, pour aller d'un état à un autre, on se déplacera uniquement sur les arêtes de l'nypercube (jamais sur les diagonales). Cela restreint le nombre de transition.

Il y a deux types de transition : celles vers le haut ("upward") et celles vers le bas ("downward"). Les transitions vers le haut se produisent lorsqu'un appel survient. Le passage d'un état à un autre par ce type de transition entraîne une augmentation de un du nombre d'unités occupées. Larson [6] explique que l'obtention de ces taux est reliée à la politique de répartition ("fixed preference") des unités. Il montre comment les calculer. Les transitions vers le bas surviennent lors d'une fin de service. Le passage d'un état à un autre par ce type de transition entraîne une diminution de un du nombre d'unités occupées. Dans le cas où tous les serveurs ont un même temps moyen de service ($1/\mu$), ce taux est de μ (voir [6]).

4.2.4 Obtention des équations à l'état d'équilibre et leur solution

À partir des taux de transition, de même que des probabilités associées aux états du modèle agrégé M/M/n, on peut obtenir un système de 2^n équations linéaires. Ce système, une fois solutionné, donne les probabilités à l'état d'équilibre des 2^n états du modèle hypercube. Larson [6]

explique comment résoudre efficacement ce système d'équations d'une manière exacte. De plus, il montre comment on peut, à partir des 2^n probabilités obtenues, calculer toutes les statistiques voulues à l'état d'équilibre.

Toutefois, si le nombre d'unités est grand ($n > 15$), alors le nombre d'équations à résoudre sera trop élevé ($2^n > 32768$). Il existe une procédure qui permet d'approximer les n charges de travail des unités en résolvant un système de n équations non-linéaires. À partir de ces approximations, Larson montre comment on peut obtenir des statistiques pour les performances à l'état d'équilibre. En ce qui concerne la qualité de ces approximations, il affirme [6] qu'elle s'améliore lorsque n grandit pour $n \leq 15$. En fait, l'erreur relative semble rarement dépasser 1 ou 2% pour $n \leq 15$. Il sera intéressant d'examiner cette affirmation pour $n > 15$.

4.2.5 Robustesse du modèle

En ce qui nous concerne, l'application du modèle se fera en considérant des situations simplifiées par rapport à la réalité. Entre autres, dans ces situations simplifiées,

i) il n'y a pas de priorité pour les appels,
ii) une seule unité est affectée pour chaque appel,
iii) il n'y a pas de relocalisation
et iv) les taux d'appel sont constants et l'état d'équilibre est supposé atteint.

Même en simplifiant de la sorte, on ne pourra se conformer exactement aux hypothèses du modèle. Comme nous le verrons au chapitre 5, l'hypothèse sur les temps de service (issus d'une loi exponentielle) n'est pas vérifiée. Cependant, Jarvis [20] affirme que le modèle est robuste à des temps de service issus d'une fonction de répartition qui s'écarte raisonnablement de celle d'une exponentielle. Cela sera à vérifier.

Larson, pour se prononcer sur la qualité de sa méthode d'approximation, n'a eu qu'à la comparer à la méthode exacte de solution du modèle hypercube, car le faible nombre d'ambulances ($n < 15$) permettait de résoudre exactement les équations à l'état d'équilibre. Le nombre d'ambulances simultanément en service à Montréal est d'au moins 26 et peut même atteindre 50 (seulement pour le service 4242) aux heures de pointe. Il n'est donc pas question de résoudre exactement les équations à l'état d'équilibre associées à ces problèmes pour obtenir des valeurs exactes des différents paramètres de l'hypercube. Nous aurons donc recours à la simulation pour nous fournir les résultats et ainsi vérifier si la précision des approximations de Larson tient toujours pour un n assez élevé. De plus, lorsque nous nous éloignerons des hypothèses de l'hypercube, la simulation sera un outil irremplaçable pour générer les paramètres associés à ces nouveaux contextes et juger si les résultats de la méthode approximative de Larson peuvent être quand même utilisables.

4.3 Méthodologie

L'utilisation du modèle hypercube suppose préalablement une analyse statistique. Celle-ci s'avère nécessaire pour vérifier les hypothèses sur le phénomène d'arrivée (de Poisson) et sur les temps de service (issus d'une loi exponentielle). Une telle analyse permettra en outre d'obtenir une description générale de la réalité. L'analyse statistique à partir des données d'Urgences Santé fera l'objet du chapitre 5.

On s'intéressera ensuite à l'évaluation de la qualité des performances calculées analytiquement à l'aide de la méthode d'approximation du modèle hypercube (car $n > 15$). C'est en comparant ces dernières avec celles obtenues par un programme de simulation qu'on y parviendra. Le programme de simulation que nous utiliserons a été développé au C.R.T. Il reproduit les situations simplifiées dont nous avons discuté en 4.3.5. Les

paramètres d'entrée (taux d'arrivée, temps de service moyen, partition du territoire, distances entre les secteurs, politique "fixed preference" seront les mêmes dans le programme de simulation que ceux utilisés avec le modèle analytique.

Cette comparaison des performances se fera graduellement à partir de deux situations. Dans la première, on étudiera la précision des charges de travail ρ_i i=1,...,n, obtenues avec la procédure d'approximation pour nombre d'unités n supérieur à vingt-cinq. La situation simulée sera, dans ce cas, exactement la même que celle du modèle hypercube. En particulier, les temps de service seront issus d'une loi exponentielle.

Dans la deuxième comparaison, on se penchera sur la robustesse des approximations des mesures (temps moyen de déplacement, charges de travail) relativement à l'hypothèse sur les temps de service. On simulera des temps de service issus de la fonction de répartition empirique obtenue à partir des données.

Nous croyons important d'accorder une bonne place à l'analyse statistique. Celle-ci s'avère nécessaire au départ afin d'obtenir correctement ce que l'on doit fournir en entrée au programme de simulation et au modèle hypercube. En particulier, les fonctions de répartition à partir desquelles les événements seront générés devront être bien identifiés; les différents paramètres devront également être bien estimés.

L'analyse statistique ne s'arrête pas là. On devra bien interpréter les résultats obtenus. Pour chaque simulation, cela veut dire qu'on devra chercher à calculer les meilleures estimations possibles des vraies performances inhérentes au contexte simulé. On apportera donc un soin particulier à la construction d'intervalles de confiance pour ces performances. De cette façon, on pourra se prononcer avec un niveau de confiance connu sur la qualité des mesures approximées par le modèle de Larson. Cette analyse "post-simulation" fera l'objet du chapitre 6.

CHAPITRE V

ANALYSE DE DONNÉES

5.1 Introduction

Une analyse de données s'impose afin d'obtenir une bonne description des différentes caractéristiques du système. Une telle analyse s'avère aussi nécessaire dans notre cas pour vérifier les hypothèses statistiques du modèle hypercube (arrivées de Poisson, temps de service issus d'une loi exponentielle).

Sur le plan descriptif, nous allons surtout étudier le phénomène de l'arrivée des demandes (ou appels). Il importe qu'il en soit ainsi, étant donné que l'analyse et la recherche de stratégies de déploiement nécessitent de bonnes estimations des taux d'appel. Il est évident que nous n'avons pas affaire à un phénomène constant. La fréquence des appels varie en fonction de plusieurs facteurs comme le temps, l'espace et le niveau de priorité. Une seule série d'estimations ne peut pas à elle seule refléter l'effet de chacun de ces facteurs, ainsi que la manière dont ils interagissent. On devra donc chercher une méthode pour obtenir plusieurs séries distinctes d'estimations, chacune tenant compte des diverses façons dont le phénomène se présente.

En plus d'analyser l'arrivée des demandes, nous examinerons les différents temps (de réponse et de déplacement), et surtout, nous vérifierons les hypothèses statistiques du modèle de Larson. Pour faire cela, des informations assez précises sont indispensables. Nous nous attarderons dans un premier temps à discuter de ces informations, de même que de la validité des données que nous avons recueillies.

5.2 Données

5.2.1 Informations nécessaires

L'analyse de l'arrivée des demandes suppose que l'on doive être capable de caractériser les appels suivant le temps, l'espace et la priorité. Pour chaque appel, les informations suivantes seront indispensables:

i) l'heure exacte de l'arrivée de l'appel,

ii) la localisation précise sur le territoire,

iii) son niveau de priorité.

Pour étudier les performances du système, on devra connaître, pour chaque unité d'urgence envoyée sur les lieux d'un appel, l'heure exacte à laquelle est survenue chacune des étapes du processus d'intervention. On devra aussi pouvoir identifier les cas où l'unité d'urgence n'a pas complété son service (annulation).

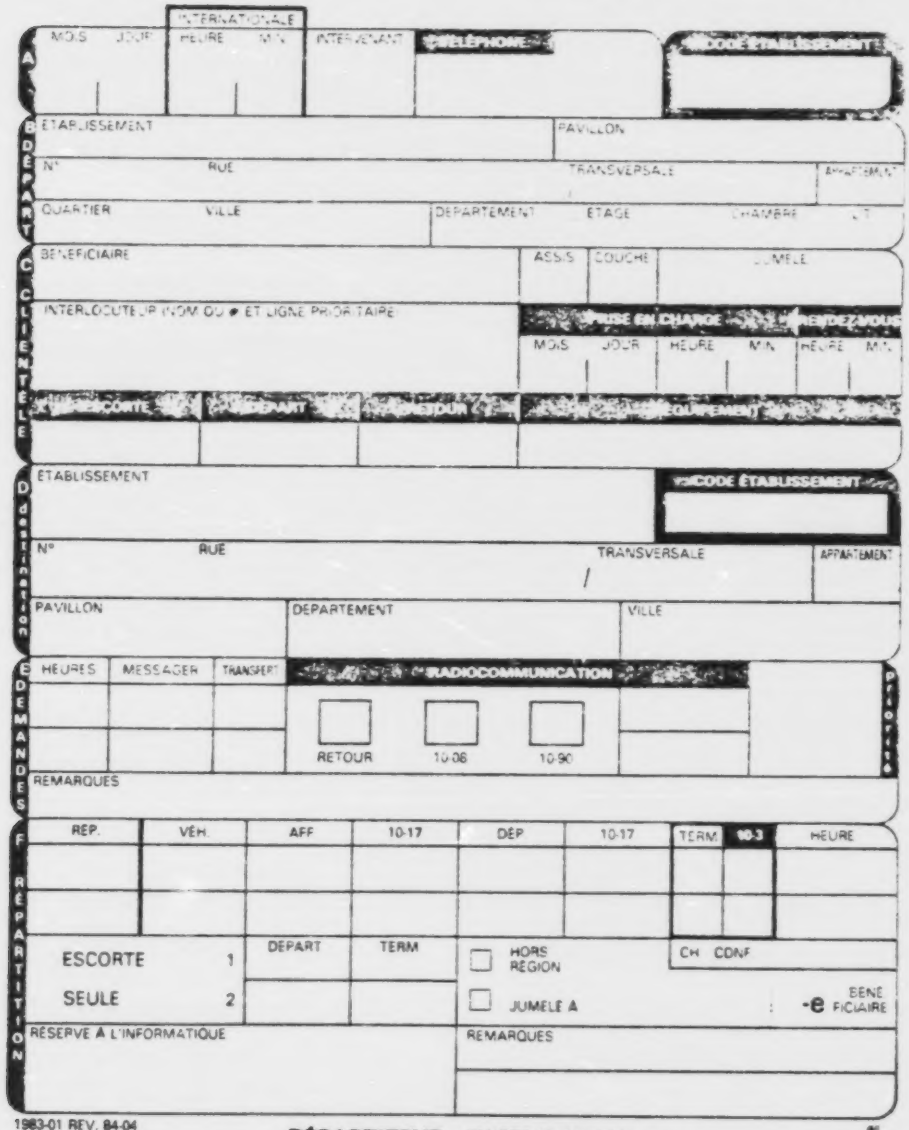
Une annulation de service peut survenir avant ou après l'arrivée de l'ambulance sur les lieux. Dans le premier cas, c'est qu'on a décidé d'envoyer une autre unité (soit qu'elle est plus près, soit qu'elle dispose d'un équipement plus approprié) ou tout simplement que l'appel comme tel est annulé. Si l'annulation survient après l'arrivée de l'ambulance sur les lieux, alors c'est qu'on a évalué sur place que le service se terminait là (faux appels, patient n'ayant pas besoin de se faire transporter).

5.2.2 Collecte de données

Urgences Santé cueille des informations pour des fins administratives et de planification. Lors de chaque appel, ces informations sont notées sur une fiche (fiche-appel). Cette dernière contient un ensemble de données prises à partir du moment où l'appel survient jusqu'au retour en disponibilité des unités affectées (voir figure 1). Tout ce qui nous intéresse se trouve sur cette fiche.

FICHE-APPEL URGENCES SANTÉ
MODULE DE L'INTER-ÉTABLISSEMENT

Z-02762 - 498



1983-01 REV. 84-04

RÉPARTITEUR - FACTURATION

FIGURE 1

Nous avons obtenu sur bandes magnétiques des données pour le mois de juin 1984. Nous n'avons sélectionné que les appels d'urgence de type 4242 pour lesquels au moins une ambulance (annulée ou non) était envoyée. En tout, nous avons constitué un fichier de 9636 cas. Pour certains appels, il pouvait y avoir plusieurs ambulances affectées. Seule celle avec le meilleur temps de réponse apparaît dans notre fichier.

5.2.3 Description des informations recueillies

- Date de la journée où l'appel est survenu (en jours/mois/an)
- Heure exacte de l'arrivée de l'appel (en heures-minutes)
- Priorité de l'appel (trois niveaux : s - 2 -4)
- Localisation de l'appel : adresse et/ou coin de rue le plus près.
- Les différentes heures dans le processus d'intervention.
  TH1 : heure lors de l'affectation de l'unité d'urgence.
  TH2 : heure lors de l'arrivée sur les lieux
  TH3 : heure lors du départ des lieux
  TH4 : heure lors de l'arrivée à l'hôpital
  TH5 : heure lors du retour en disponibilité.
- Code spécial
  ARR : permet de détecter si l'ambulance a été annulée (avant ou après son arrivée). ARR = 1003 → Ambulance annulée.

5.2.4 Validité des données

L'identification du lieu exact de l'appel fut ce qui nous a causé le plus de problèmes. Nous avons d'abord procédé d'une façon automatique à partir de l'adresse en utilisant les "Area Master Files" (AMF) fournies par Statistiques Canada [23].

Les AMF consistent en un ensemble de fichiers contenant diverses données géographiques servant à décrire le réseau des voies carossables et

autres sortes de courbes dans une ou plusieurs régions données. Il est donc possible à partir de ces informations de localiser des appels de façon précise sur un réseau.

Pour pouvoir utiliser les AMF, l'adresse doit être très bien codée. Or, dans les données que nous avions à notre disposition, les adresses n'étaient pas codées de façon standard (c'est-à-dire avec un champ de caractère bien défini, orthographe correcte, nom complet). Ainsi, toutes les adresses ont dû subir des transformations. Malgré cela, près de 1 500 appels n'ont pu être localisés, l'adresse étant trop incomplète. Nous avons dû aller consulter les fiche-appels sur place afin de recueillir correctement les informations. À la fin de cet exercice, tous les appels ont pu être localisés.

Pour ce qui est de la validité des autres données obtenues, il y avait 3.7% de cas où la priorité était manquante. Pour le reste, cette proportion était inférieure à 1%. Donc les informations que nous avons à notre disposition peuvent être considérées comme étant valables.

5.3 Analyse de la demande

5.3.1 Buts

Le but de notre analyse de la demande consiste à identifier les facteurs influençant la variabilité du phénomène et à catégoriser les appels en fonction de ces facteurs. De cette façon, on pourra obtenir plusieurs séries distinctes d'estimation des taux d'appels, chacune d'entre elles correspondant à une façon dont le phénomène se présente.

Il est à noter que cette analyse ne pourra pas réellement servir à faire des prédictions pour évaluer les taux d'appels futurs. La construction d'un modèle de prédiction nécessite des données rendant compte du phénomène sur plusieurs années.

Les appels peuvent être caractérisés par plusieurs facteurs. Essentiellement, nous allons nous intéresser à trois d'entre eux qui nous semblent les plus importants et pour lesquels nous avons des informations. Ces facteurs sont : le temps, l'espace et le niveau de priorité.

5.3.2 Visualisation des données

Dans un premier temps, regardons comment les appels sont répartis pour chacun des facteurs. Pour ce qui est de la répartition des appels dans l'espace, nous comptons utiliser pour le modèle hypercube, une partition assez fine du territoire. Pour ce faire, nous avons regroupé les appels par secteur de recensement (468 secteurs pour le territoire considéré - Réf. [23]). Nous aurons besoin d'évaluer les taux d'appels dans chacun de ces secteurs. Toutefois, il est difficile de visualiser la répartition des appels sur le territoire pour une partition aussi fine. Nous avons constitué 26 régions et dénombré les appels pour chacune d'elles. Ce découpage correspond à celui utilisé par Urgences Santé pour ses activités de planification (voir Figure 2). Urgences Santé a prédéfini 26 postes d'attentes prioritaires. À partir de ces 26 localisations, nous avons défini nous-mêmes les 26 régions.

La figure 3 montre la fréquence des appels selon le niveau de priorité. Il en existe trois. Ceux-ci sont reliés à une borne maximale sur le temps de réponse. On notera que la majorité des appels (6564 sur 9636) sont de première priorité (S).

Les taux d'appel peuvent facilement être estimés selon leur provenance (secteur de recensement) et leur niveau de priorité. Cependant, nous considérons une telle série d'estimations pour une période où le phénomène est constant par rapport au temps. Si l'on regarde le tableau 1 on remarque que le nombre d'appels fluctue d'une journée à l'autre. Il en est de même à l'intérieur d'une même journée. On se doit d'étudier ces fluctua-

Note : $y^{\hat{}}(z)$

x : no. du poste d'attente
y : # d'appels
z : % d'appels

Répartition des appels
sur le territoire
Juin 1984

FIGURE 2

NIVEAU DE PRIORITÉ (code)	BORNE SUPÉRIEURE SUR LE TEMPS DE RÉPONSE	FRÉQUENCES DES APPELS
S	5 minutes	6564 (68.1%)
2	20 minutes	2146 (22.3%)
4	≈60 minutes	568 (5.9%)
Manquant		358 (3.7%)

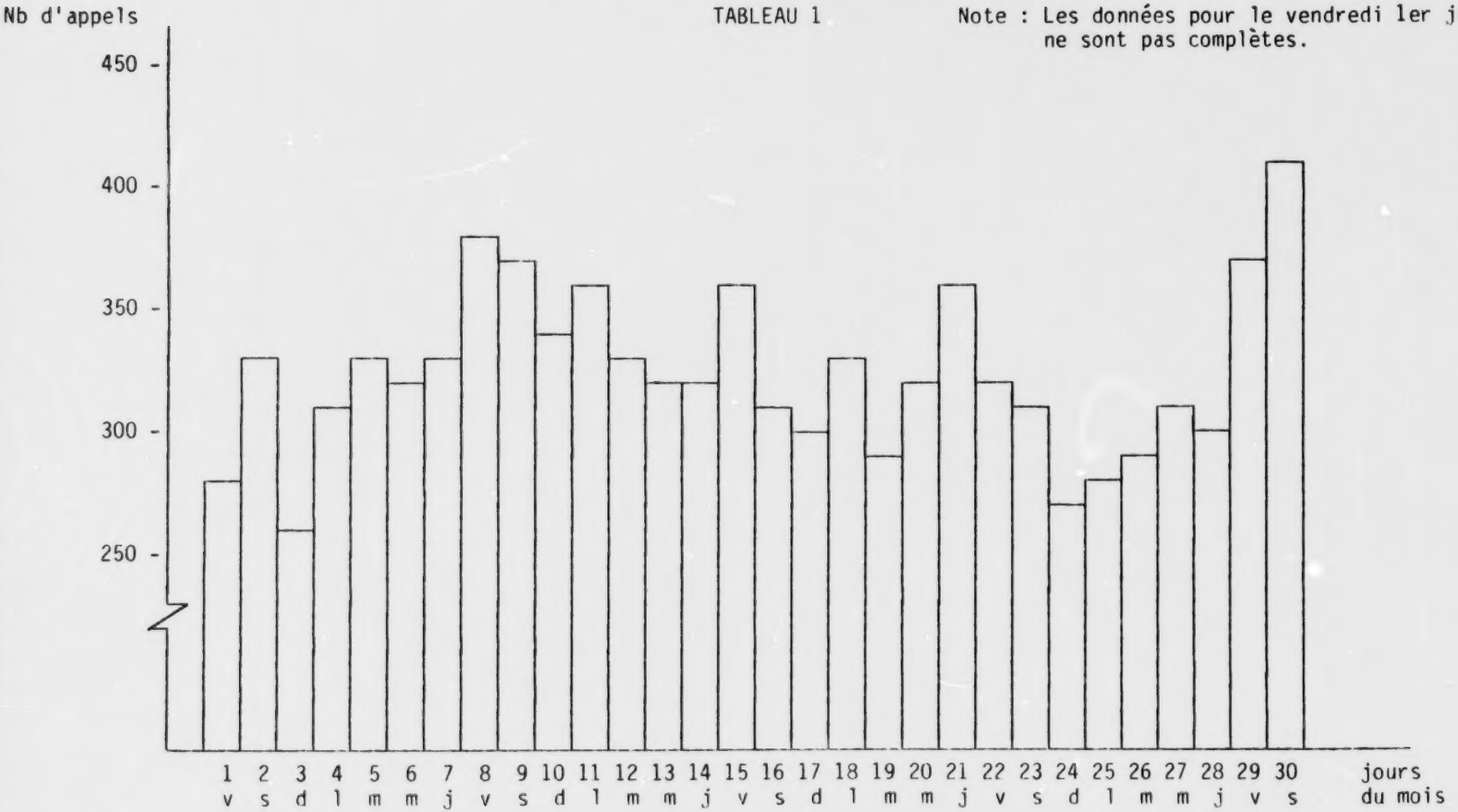
FRÉQUENCE DES APPELS
PAR NIVEAU DE PRIORITÉ

FIGURE 3

RÉPARTITION DES APPELS PAR JOUR POUR LE MOIS DE JUIN

TABLEAU 1

Note : Les données pour le vendredi 1er juin ne sont pas complètes.



tions dans le temps afin de catégoriser les appels par rapport à ce facteur. De cette façon, on pourra déterminer des périodes où le phénomène d'arrivées est constant dans le temps, et ainsi, obtenir des estimations cohérentes des taux d'appel.

5.3.3 Étude de la demande en fonction du temps

5.3.3.1 Identification des facteurs de variabilité

La demande varie dans le temps en fonction de trois facteurs qui sont : l'aspect saisonnier, la période de la semaine et la période de la journée.

L'aspect saisonnier : La demande diffère d'une saison à l'autre pour toutes sortes de raisons. Cela peut dépendre du climat, des conditions routières, etc. Dans le cas du mois de juin, on peut penser a priori que le début de la saison estivale peut influencer la demande, surtout après le 24 juin (voir Tableau 1). Nous voulons, ici, identifier la portion du mois où l'aspect saisonnier n'a pas d'effet sur la variabilité de la demande.

La période de la semaine : À l'intérieur d'une même semaine, le phénomène se présente de différentes façons (par exemple : fin de semaine vs semaine). On cherche ici à regrouper ensemble les jours de la semaine pour lesquels la demande est répartie de la même façon. On aura alors différents types de journées.

La période de la journée : Pour chaque type de journée que l'on aura identifié, nous voudrions déterminer les périodes de pointe en regroupant les heures où le taux d'appel est homogène et évidemment le plus élevé. Nous nous limitons à l'étude des périodes de pointe, car ce sont les pires cas et que nous testerons sur celles-ci si le phénomène d'arrivée est un processus de Poisson.

5.3.3.2 Méthodologie et résultats

Nous caractériserons le phénomène de l'arrivée de la demande par le nombre d'appels par heure.

5.3.3.2.1 Identification de la portion du mois où le facteur saisonnier n'as pas d'effet

Les données pour les jours du 1er et du 30 juin n'étant pas complètes, nous avons sélectionné les jours du 2 au 29 juin inclusivement. Une période où l'effet saisonnier n'aurait pas d'effet serait telle que la répartition des appels par heure serait identique pour chaque lundi; pour chaque mardi; etc. Nous allons faire une étude par tableau de contingences et utiliser le test du CHI-2.

Par exemple, pour une période donnée, nous voudrions savoir si les appels sont répartis de la même façon pour les quatre lundi de cette période. Nous construisons le tableau de contingences suivant :

Jour / Heure	Jour 1	Jour 2	Jour 3	Jour 4	TOTAL
0:00- 0:59	M_{11}	M_{12}	M_{13}	M_{14}	M_{1+}
1:00- 1:59	M_{21}	M_{22}	M_{23}	M_{24}	M_{2+}
2:00- 2:59	M_{31}	M_{32}	M_{33}	M_{33}	M_{3+}
.					
.					
.					
21:00-21:59	M_{221}	M_{222}	M_{223}	M_{224}	M_{22+}
22:00-22:59	M_{231}	M_{232}	M_{233}	M_{234}	M_{23+}
23:00-23:59	M_{241}	M_{242}	M_{243}	M_{244}	M_{24+}
Total	M+1	M+2	M+3	M+4	M

Notons M_{ij} = nombre d'appels qui se sont produits la $i^{ème}$ heure du $j^{ème}$ lundi

$M_{i+} = \sum_j M_{ij}$, $M_{+j} = \sum_i M_{ij}$, $M = \sum_i \sum_j M_{ij}$

P_{ij} = probabilité qu'un appel survienne à la $i^{ème}$ heure du $j^{ème}$ lundi

P_i = probabilité qu'un appel survienne à la $i^{ème}$ heure

P_j = probabilité qu'un appel survienne le $j^{ème}$ lundi

Le test

On teste l'hypothèse H_0 : heure de l'appel indépendante du jour ($p_{ij} = p_i p_j$). La statistique utilisée pour le test est $S = \sum_i \sum_j \frac{M_{ij} - M_{i+} M_{+j}/n)^2}{(M_{i+} M_{+j}/n)}$. Dans Lindgren [25, p. 438], on démontre que la région critique du test est $S<k$ et dans Cramer [26], on démontre que sous H_0, S est approximativement X_2 avec (24-1)(4-1) degrés de liberté. On fixe α, le niveau du test, et on calcule X^2 observé. Si X^2 observé $> X^2$ critique, on rejette H_0.

Résultats

En faisant des tests X^2 au niveau 5% pour chacune des journées de la semaine, nous avons pu isoler la période du 3 au 23 juin où il n'y avait pas d'effet saisonnier, soit une période de trois semaines. Le tableau 2 montre les résultats des tests. Le nombre de degrés de liberté est de 46 ((24-1) x (3-2)). La valeur X^2 critique est égale à 62.49.

Hypothèse d'indépendance pour	X^2_o	Décision
Heure-appel x Lundis	47.80	Non-rejet de Ho
Heure-appel x Mardis	46.85	Non-rejet de Ho
Heure-appel x Mercredis	40.55	Non-rejet de Ho
Heure-appel x Jeudis	48.70	Non-rejet de Ho
Heure-appel x Vendredis	57.39	Non-rejet de Ho
Heure-appel x Samedis	55.19	Non-rejet de Ho
Heure-appel x Dimanches	50.68	Non-rejet de Ho

RÉSULTATS DES TESTS CHI-2

TABLEAU 2

5.3.3.2.2 Identification des types de journées

On veut déterminer les journées de la semaine pour lesquelles le nombre d'appels par heure est réparti de la même façon pour la période du 3 au 23 juin. Encore ici, on utilise le test du CHI-2.

Par exemple, on veut savoir si la demande est répartie de la même façon les samedis et les dimanches. On construit le tableau de contingences suivant :

Heure \ Jour	Samedi 1	Dimanche 1	Samedi 2	Dimanche 2	Samedi 3	Dimanche 3
0:00- 0:59						
1:00- 1:59						
.						
.						
.						
21:00-21:59						
22:00-22:59						
23:00-23:59						

On teste Ho : heure de l'appel indépendante du jour. Si Ho n'est pas rejetée, on peut alors supposer que la proportion d'appels par heure est répartie de la même façon les samedis et les dimanches.

Résultats

Nous avons trouvé trois types de journée :

- fin de semaine (samedi et dimanche)
- semaine (du lundi au jeudi)
- vendredi

Les résultats se trouvent au tableau 3.

Type de journée	# de degrés de liberté	X_o^2	X_c^2	Décision
Fin de semaine	115	127.67	140.66	Non-rejet de Ho
Lundi-jeudi	253	259.09	290.70	Non-rejet de Ho
Vendredi	46	57.39	62.49	Non-rejet de Ho

Note : Ho : Heure de l'appel indépendante du jour.

RÉSULTATS DES TESTS POUR DÉTERMINER LES TYPES DE JOURNÉES

TABLEAU 3

Avant de continuer notre étude pour déterminer les périodes de pointe de chaque type de journées, nous allons examiner un peu les statistiques. Le tableau 4 indique le nombre d'appels par jour pour chacune des semaines du dimanche 3 juin au samedi 23 juin. Les tests que nous avons faits nous ont permis de regrouper les jours pour lesquels le nombre d'appels était réparti de la même façon (lundi-jeudi, samedi-dimanche, vendredi). Cela ne veut pas dire que le nombre d'appels est homogène pour ces journées comme en fait foi le tableau 4. Toutefois, on peut constater que la fréquence des appels est moins intense la fin de semaine et que le vendredi semble la journée la plus occupée.

	1ère semaine	2e semaine	3e semaine	TOTAL
Dimanche	259	345	284	888
Lundi	305	360	324	989
Mardi	323	327	297	947
Mercredi	322	318	319	959
Jeudi	333	319	355	1007
Vendredi	373	353	318	1044
Samedi	363	304	303	970
TOTAL	2278	2326	2200	6804

RÉPARTITION DU NOMBRE D'APPELS PAR JOURS
ET PAR SEMAINES DU 03/06 au 23/06

TABLEAU 4

5.3.3.2.3 Identification des périodes de pointe

Pour chaque regroupement de jours déterminé à la section précédente, on cherche maintenant à regrouper des heures pour lesquelles le nombre d'arrivées observées est homogène afin de déterminer des périodes de pointe. Pour chaque type de journée, on identifie d'abord un groupe d'heures consécutives durant lesquelles les fréquences des appels sont les plus élevées. On teste ensuite l'homogénéité du nombre d'appels pour chacune de ces heures. On utilise un test CHI-2 sur des proportions.

Le test

Soient k heures consécutives. On définit

O_i = nombre d'arrivées observées durant l'heure i, i=1,2,...,k.

$M = \sum_{i=1}^{k} Oi$ = nombre total d'arrivées observées durant la période

$P_i = \frac{O_i}{n}$ = proportion d'arrivées observées durant l'heure i.

On teste Ho : $p_1 = p_2 = \ldots = p_k = \frac{1}{k}$

Soient $\gamma_i = n \times \frac{1}{k}$ =la fréquence théorique pour l'heure i, i=1,2,...,k.

La statistique utilisée est $S = \sum_{1}^{k} \frac{(O_i - \gamma_i)^2}{\gamma_i}$. On démontre que la région critique est de la forme S > K et que sous Ho, S est approximativement X^2 avec k-1 degrés de liberté(1). En fixant α, le niveau du test, on détermine k_α . Si S observé > k_α on rejette Ho.

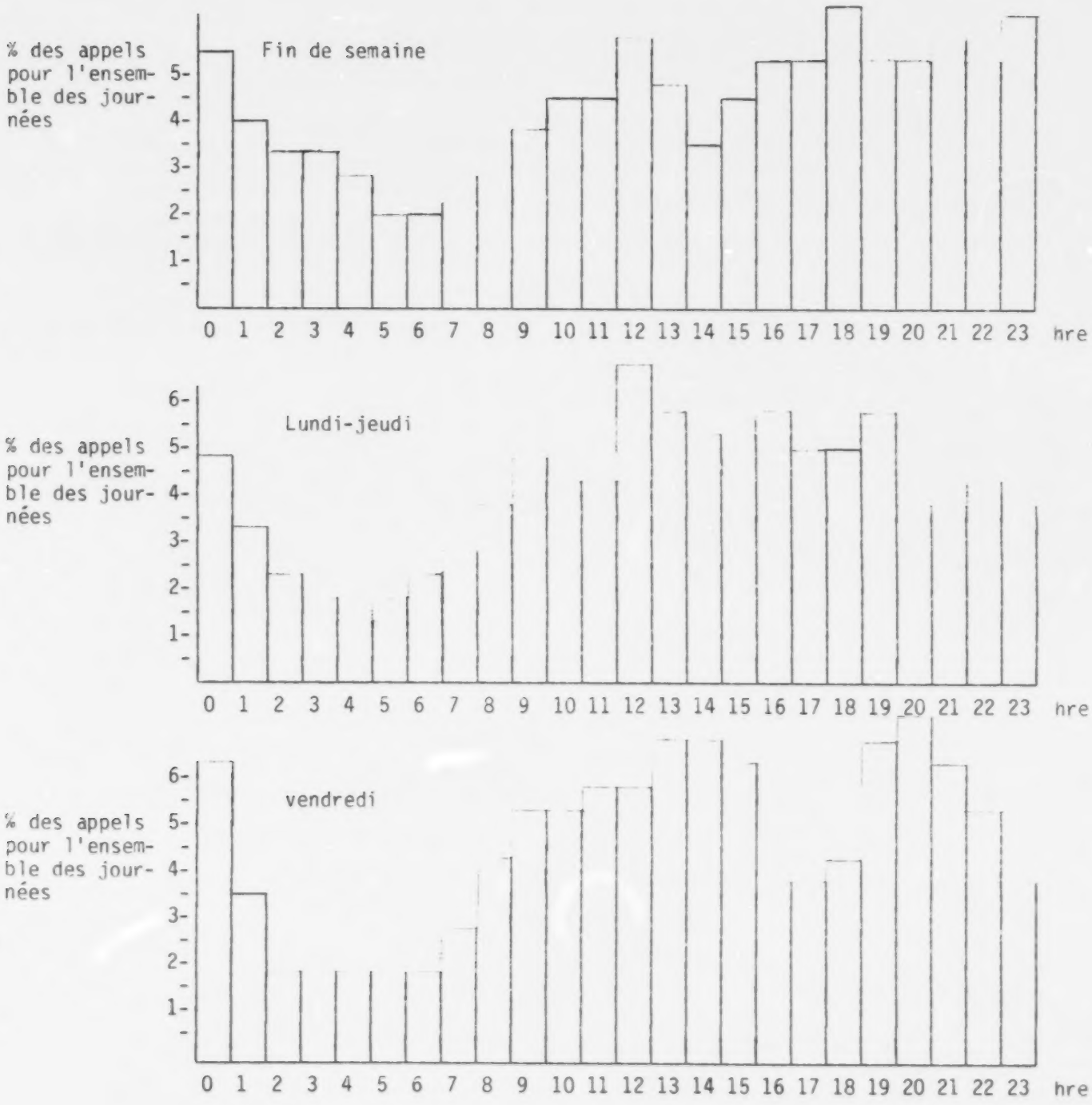
Résultats

Le tableau 5 indique la répartition du nombre d'appels cumulatifs par heure pour chacun des types de journée. Nous avons additionné ensemble les fréquences des appels pour toutes les journées appartenant à un même type. Par exemple, pour la fin de semaine, nous avons cumulé les fréquences pour tous les samedis et les dimanches. Les histogrammes du tableau 5 nous ont permis de déterminer a priori les périodes de pointe. Nous avons ensuite testé au niveau 5% l'homogénéité du nombre cumulatif d'appels pour chacune de ces périodes. Le tableau 6 indique les résultats que nous avons obtenus. Les périodes de pointe sont :

---

(1) Voir Cramer [26]

TABLEAU 5


% des appels pour l'ensemble des journées
Fin de semaine
1
2
3
4
5
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 hre
% des appels pour l'ensemble des journées
Lundi-jeudi
1
2
3
4
5
6
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 hre
% des appels pour l'ensemble des journées
vendredi
1
2
3
4
5
6
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 hre

- de 16 à 23 heures pour la fin de semaine
- de 12 à 16 heures pour la période lundi-jeudi
- de 10 à 15 heures et de 19 à 22 heures pour le vendredi.

Il est à remarquer que, pour la présente série de tests, nous travaillons avec l'hypothèse selon laquelle les journées de chaque type peuvent être regroupées. Nous avons testé cette hypothèse précédemment au niveau 5%. Ce qui veut dire que les tests que nous venons de faire sont correlés avec les précédents. Ainsi, la probabilité de faire une erreur de premier type (rejeter Ho si Ho est vraie) est ici plus grande que ne le laisse voir les résultats. Toutefois, les statistiques X^2 calculées étant peu élevées dans trois cas sur quatre (tableau 6), nous allons quand même utiliser les décisions qui en découlent (non rejet de Ho).

Type de journée	Période de pointe	Hypothèse Ho	#degrés de lib.	X^2_c	X^2_o	décision
Fin de semaine	16-23h	$p_i = \frac{1}{8}$	7	14.1	2.31	non-rejet de Ho
Lundi-Jeudi	12-16h	$p_i = \frac{1}{5}$	4	9.49	7.16	non-rejet de Ho
Vendredi	10-15h	$p_i = \frac{1}{6}$	5	11.1	3.46	non-rejet de Ho
	19-22h	$p_i = \frac{1}{4}$	3	7.81	5.53	non-rejet de Ho

RÉSULTATS DES TESTS D'HOMOGÉNÉITÉ DU NOMBRE D'APPELS

TABLEAU 6

5.3.4 Conclusion

L'étude du facteur temps est maintenant terminée. Nous pouvons donc catégoriser les appels par rapport à trois types de journées : fin de

semaine, lundi-jeudi et vendredi. Nous n'étudierons que les périodes de pointe de chacun de ces types. Nous calculerons des estimations des taux d'appel pour ces périodes et nous testerons si le phénomène d'arrivées est Poisson sur chacune d'elles.

5.4 Vérification des hypothèses statistiques du modèle hypercube

5.4.1 Arrivées des demandes Poisson

5.4.1.1 Formulation de l'hypothèse

D'un point de vue statistique, l'utilisation du modèle hypercube sous-entend deux hypothèses. La première concerne le phénomène d'arrivée et s'énonce comme suit :

> Étant donné la partition du territoire en N atomes, la répartition de l'arrivée des demandes dans chaque atome i, i=1,...,N. doit être celle d'un processus de Poisson avec paramètre λ_i.

Un tel processus de Poisson est un processus constant dans le temps. On doit donc chercher à faire des tests pour des périodes de temps où le phénomène est à peu près constant. Nous allons utiliser les périodes de pointe déterminées précédemment.

Nous ferons des tests globaux, c'est-à-dire pour l'ensemble du territoire. Il nous sera par contre impossible de tester localement, c'est-à-dire par région. Comme nous le verrons, le petit nombre d'observations, de même que les méthodes utilisées ne le permettent pas.

5.4.1.2 Méthodologie

Il existe deux façons de tester si le phénomène d'arrivée est Poisson sur une période de temps. La première consiste à observer le nombre d'occurrences du phénomène par intervalles de temps $[0,t]$ sur la période.

On dispose alors d'autant d'observations qu'il y a d'intervalles $[0,t]$ dans ladite période. On tente ensuite d'ajuster la répartition empirique de ces observations à celle d'une loi de Poisson à l'aide d'un test CHI-2.

L'autre façon consiste à calculer les temps entre chacune des arrivées survenant durant la période. On démontre que si le phénomène est un processus de Poisson (λ), alors la répartition des temps inter-arrivées est celle d'une loi exponentielle de moyenne $1/\lambda$. On n'a alors qu'à ajuster la fonction de répartition des temps observés à celle d'une exponentielle ($1/\lambda$).

Nous avons d'abord utilisé la méthode des temps inter-arrivées, celle-ci nous apparaissant la plus puissante. En effet, dans ce cas on peut compter sur un plus grand nombre de données. De plus, pour l'ajustement, on utilise un test de type Kolmogorov (test de Lilliefors). Ce dernier est plus puissant qu'un test CHI-2. Malheureusement, cette méthode n'a pu fonctionner et nous allons voir pourquoi.

Nous avons calculé les temps inter-arrivées pour chacune des journées de fin de semaine. Il y avait six journées. Ensuite, pour chacune de ces journées, nous avons testé si les temps étaient répartis suivant une loi exponentielle. Nous n'avons pas regroupé les journées ensemble, car il n'est pas évident que les fonctions de répartition étaient homogènes d'une journée à l'autre.

Le test de Lilliefors (Ref [27] p. 362)

Soient $x_1, x_2, x_3, \ldots, x_n$ les temps inter-arrivées ordonnés ($x_1 \leq x_2 \leq \ldots \leq x_n$).

On définit d'abord la fonction de répartition empirique comme suit :

$$F_n(x) = \begin{cases} 0 \text{ si } x < x_1 \\ k/n \text{ si } x_k \leq x < x_{k+1} \\ 1 \text{ si } x \geq n \end{cases}$$

On veut tester $Ho : F_n(x) = F(x) = 1-e^{-\lambda x}$, pour tout x

Vs $H_1 : F_n(x) \neq F(x)$ pour au moins un x. λ est estimé par $\frac{n}{\sum_1^n x_i}$ (l'estimateur à vraisemblance maximale). La statistique utilisée est $D = \sup_x |F_n(x) - F(x)|$. La région critique est de la forme $D \quad k$. Les valeurs critique k_α ont été obtenues par Lilliefors [28] à l'aide de méthodes de Monte-Carlo.

Les résultats des tests se trouvent au tableau 7. On rejette Ho dans tous les cas.

Suite à ces résultats, nous avons utilisé une méthode graphique simple. Pour chaque x_i représenté sur l'axe vertical, on associe la valeur $\log \left| \frac{m}{m-i+\frac{1}{2}} \right|$ sur l'axe horizontal. D'autre part, on trace la droite passant par l'origine et dont la pente est $1/\bar{x} = \sum_1^n x_i / n$. Si les observations proviennent d'une loi exponentielle, les coordonnées se situent sur la droite.

La figure 3 représente les résultats de cette méthode pour le premier samedi. Il est curieux de constater que les points observés soient si près de la droite. On aurait de bonnes raisons de penser que les temps inter-arrivées sont exponentiels. Or les résultats des tests ne concordent pas.

Dans les données que nous avons, le temps a été mesuré d'une manière discrète (en minutes). On a donc un échantillon aléatoire d'une variable discrète. Le test de Lilliefors ne peut être appliqué que lorsque la variable en question est continue. Autrement, la convergence de la statistique D n'est pas assurée [28]. Pour cette raison, on ne peut se fier aux résultats de ce test.

Cela est malheureux, car cette méthode nous permettait de disposer d'un plus grand nombre de données. Il aurait donc été possible de tester le phénomène localement.

Nous avons dû utiliser la première méthode, soit celle qui consiste à compter le nombre d'occurences par intervalles de 15 minutes(1) et à tester si le phénomène est Poisson à l'aide d'un test CHI-2. Un tel test nécessite que l'on regroupe des observations. Plus on dispose d'observations, plus le nombre de classes est grand et plus le test est puissant. Malheureusement, dans les tests que nous avons faits, nous n'avions jamais plus de cinq classes. On ne peut donc considérer ces tests comme étant très puissants.

Journée	Moyenne	Variance	#d'observations	Dc (5%) (10%)	Do	Décision
dimanche 3	4.20	13.09	112	0,10 0,09	0.16	On rejette Ho "
samedi 9	3.22	9.66	146	0,089 0,08	0,14	" "
dimanche 10	3.45	11.96	137	0,09 0,08	0,15	" "
samedi 16	3.84	12.27	123	0,10 0,08	0,14 0,14	" "
dimanche 17	3.53	9.51	135	0,09 0,08	0,13	" "
samedi 23	3.90	11.86	122	0,097 0,088	0,15	" "

RÉSULTATS DES TESTS DE LILLIEFORS (Journées de fin de semaine)

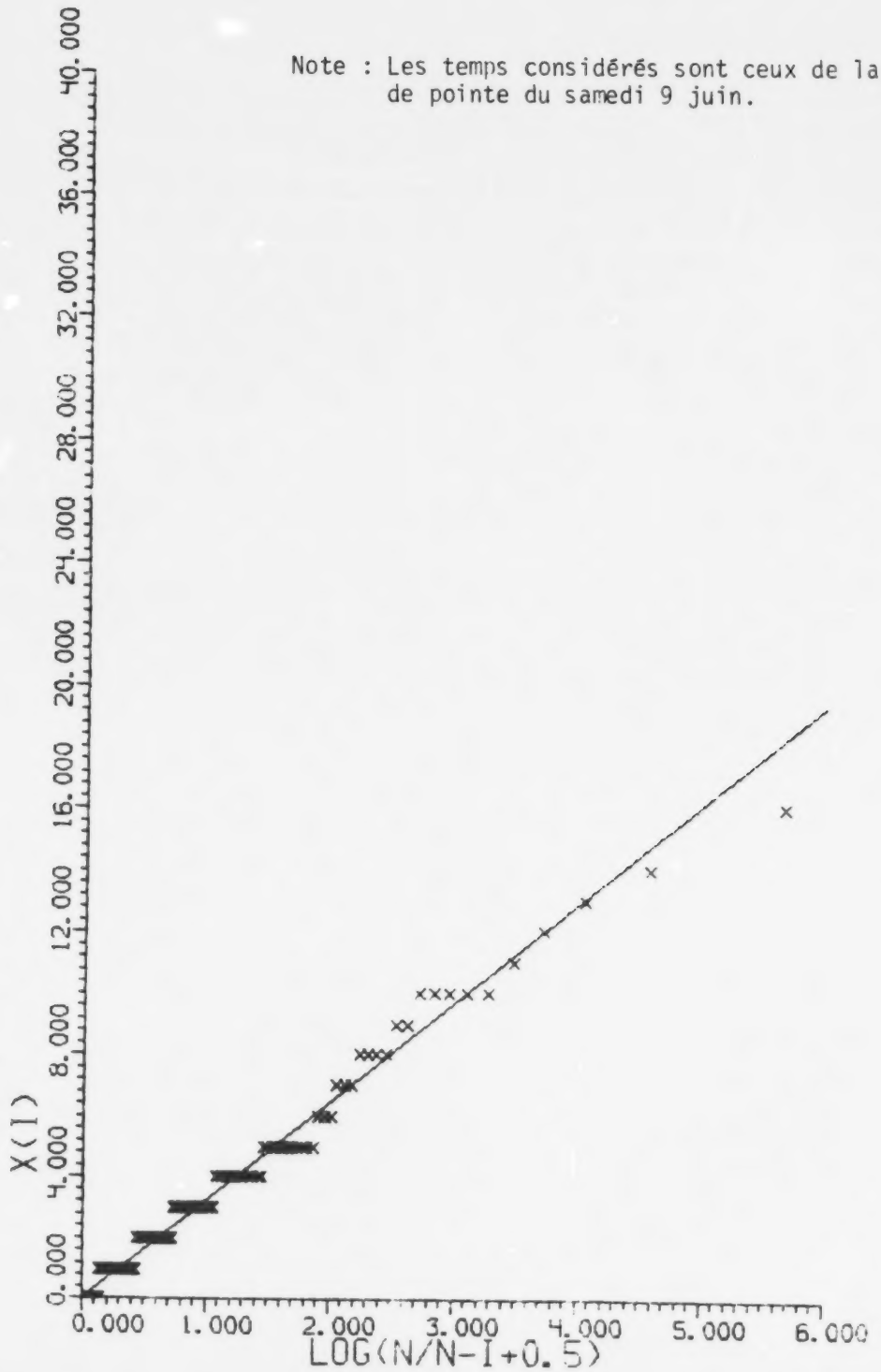
TABLEAU 7

(1) Nous avons pris des intervalles de cette longueur sous suggestion de [29] chapitre 8.

Méthode graphique de la droite
pour les temps inter-arrivées

FIGURE 3

Note : Les temps considérés sont ceux de la période de pointe du samedi 9 juin.



Le test CHI-2

Soit les classes que l'on constitue à partir des observations. Ces classes sont des intervalles ici. Soit

O_j = fréquence observée, j = 1,...,k

γ_j = fréquence théorique, j = 1,...,k.

On dispose de n observations. On peut alors obtenir γ_j par $\gamma_j = n\Pi_j$ où Π_j= Probabilité que le nombre d'occurences d'un processus de Poisson tombe dans la classe i. On voit que c'est le test CHI-2 que nous connaissons bien.

Les tableaux 8, 9 et 10 rendent compte des résultats respectivement pour les périodes de fin de semaine, de semaine (lundi-jeudi) et du vendredi. En aucun cas, on ne rejette Ho.

5.4.1.3 Conclusion

La méthode utilisée pour tester l'hypothèse sur l'arrivée des demandes est peu puissante. Par contre, elle ne conduit pas à des résultats significatifs au sens d'un rejet de l'hypothèse. En plus des tests que nous avons faits,on peut aussi examiner les moyennes et les variances des observations. Celles-ci doivent être égales pour un processus de Poisson (voir tableaux 8, 9 et 10). Enfin, la méthode graphique de la droite pourrait aussi être utilisée avec les temps inter-arrivées. Mais de cette façon, il faut prendre conscience que l'on s'éloigne de la démarche rigoureuse.

En conclusion, les moyens que nous avons ne nous permettent pas de rejeter l'hypothèse Poisson pour l'ensemble du territoire. Nous allons donc pouvoir la supposer sans contradiction apparente. De plus, ne pouvant pas tester le phénomène localement, nous supposerons que les appels arrivent suivant un processus de Poisson de moyenne λ_i dans chaque atome i.

Lundi-jeudi : Résultats des tests CHI-2 Poisson

α = 5% 20 observations par période de pointe

Jour	Moyenne	Variance	d.l	# classes avec fréq. théorique<5	X^2o	X^2c	Décision
Lundi 04/06	4.40	3.52	2	0	1,81	5,99	rejette pas Ho
Mardi 05/06	5.00	5.37	2	0	0,03	5,99	"
Mercredi 06/06	4.55	4.37	2	0	0,85	5,99	"
Jeudi 07/06	4.60	4.67	2	0	0,13	5,99	"
Lundi 11/06	4.60	3.73	2	0	0,71	5,99	"
Mardi 12/06	4.25	3.46	2	0	0,34	5,99	"
Mercredi 13/06	4.40	6.57	2	0	0,16	5,99	"
Jeudi 14/06	5.10	4.83	2	0	0,37	5,99	"
Lundi 18/06	4.55	2.36	2	0	0,16	5,99	"
Mardi 19/06	4.25	4.30	2	0	0,02	5,99	"
Mercredi 20/06	3.65	4.87	2	0	0,35	5,99	"
Jeudi 21/06	5.55	4.89	2	0	2,86	5,99	"

Nombre d'observations : 20

RÉSULTATS DES TESTS CHI-2 - Lundi au jeudi

TABLEAU 8

Fin de semaine : Résultats des tests CHI-2 Poisson

α = 5%

Jour	Moyenne	Variance	d.l	# classes avec fréq. théorique<5	X^2o	X^2c α=5%	Décision
Dimanche 03/06	3.53	3.10	3	0	2,70	7,81	On ne rejette pas Ho
Samedi 09/06	4.59	5.99	4	0	1,99	9,49	"
Dimanche 10/06	4.31	5.70	4	1	3,61	9,49	"
Samedi 16/06	3.88	3.02	4	0	5,96	9,49	"
Dimanche 17/06	4.25	4.19	3	0	1,41	7,81	"
Samedi 23/06	3.84	3.62	4	0	1,48	9,49	"

Nombre d'observations : 32

RÉSULTATS DES TESTS CHI-2 - Fin de semaine

TABLEAU 9

Résultats des tests CHI-2P Poisson (vendredi)

1) Période de pointe 10:00 - 15:00

Jour	Moyenne	Variance	d.l	# classes avec fréq. théorique<5	X^2o	X^2c	Décision
08/06	4.958	5.085	3	1	1,98	7,81	On ne rejette pas Ho
15/06	4.625	3.114	3	2(4.49,4.45)	1,77	7,81	"
22/06	4.5	7.217	3	2(4.56,4.10)	0,43	7,81	"

2) Période de pointe 19:00 - 22:00

Jour	Moyenne	Variance	d.l	# classes avec fréq. théorique<5	X^2o	X^2c	Décision
08/06	5.438	7.462	2	1(3.34)	0,48	5,99	On ne rejette pas Ho
15/06	5.625	8.117	2	0	2,99	5,99	"
22/06	5.063	2.729	2	1(4.10)	0,67	5,99	"

RÉSULTATS DES TESTS CHI-2 - Vendredi

TABLEAU 10

5.4.2 Temps de service

Les données que nous avons ne nous permettent pas de différencier chaque serveur. On se contentera donc de faire une analyse globale. Entre autres, on veut vérifier l'hypothèse selon laquelle les temps de service suivent une loi exponentielle. Il est raisonnable de penser a priori que la répartition de ces temps diffère selon la nature du service en question. Nous avons séparé les 9636 cas en deux catégories : les services annulés(1) et les services complétés. Le tableau 11 rend compte des statistiques obtenues.

Type de service	Nombre	Moyenne	Variance	Ecart-type
Annulé	2067 (21.5%)	25.31	449.53	21.20
Complété	7569 (78.5%)	50.70	390.61	19.76
Total	9636	45.25	511.88	22.62

STATISTIQUES POUR LES TEMPS DE SERVICE (en minutes)

TABLEAU 11

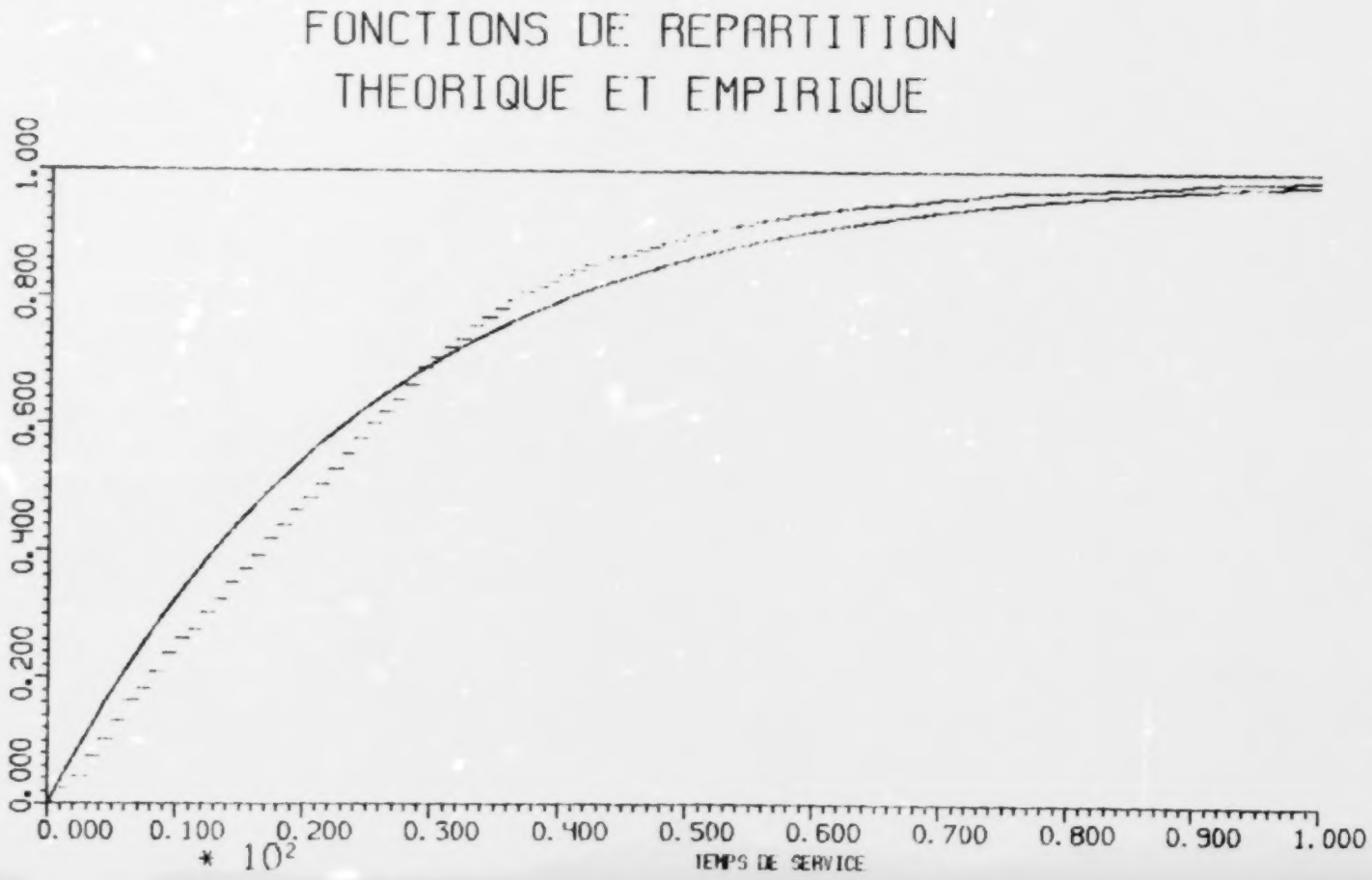
Pour chacun des deux types de service, nous avons tracé les fonctions de répartition empirique et théorique. La fonction théorique était celle de l'exponentielle dont le paramètre λ était estimé par l'inverse de la moyenne (voir figure 5 et 6). Il est clair que, pour le cas des services complétés, les temps ne sont pas exponentiels. Pour les services annulés, nous avons utilisé le test de Lilliefors décrit précédemment. La valeur critique est Dc = 0,0237 et la valeur observée Do= 0,10 (α = 5%). Ce qui implique le rejet de l'hypothèse nulle.

---

(1) Note : Appel annulé ARR = 1003

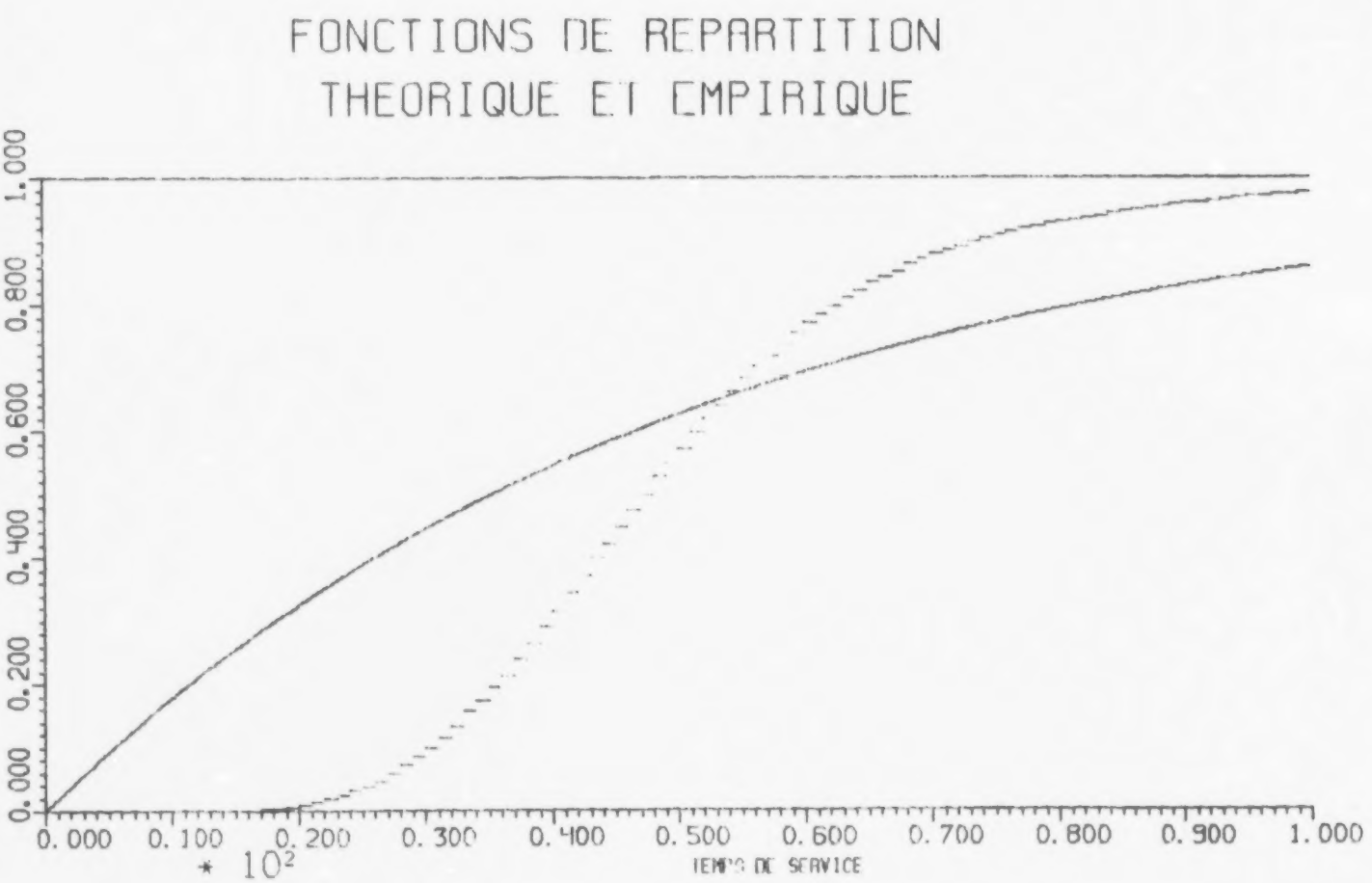
Ajustement des temps de service
de type annule à une loi exponentielle

FIGURE 5



Ajustement des temps de service
de type complete à une loi exponentielle

FIGURE 6



Il serait intéressant d'essayer d'ajuster les temps de service à une loi de type Erlang. Une telle famille de loi de probabilité englobe l'exponentielle (une Erlang de paramètres k est une somme de k exponentielles indépendantes). Nous avons fait des tests d'ajustement CHI-2 et ceux-ci se sont avérés infructueux. On peut quand même dire que les distributions empiriques se rapprochent de celles d'une Erlang de paramètres k=2 pour les services annulés et k=8 pour les autres (voir figures 7 et 8).

5.5 Temps de réponse et de déplacement

5.5.1 Temps de réponse

Le temps de réponse considéré ici est celui que nous avons défini au chapitre 2 (heure arrivée sur les lieux - heure appel). Ce n'est pas le même que celui défini par Urgences Santé (eux considèrent plutôt le temps de déplacement).

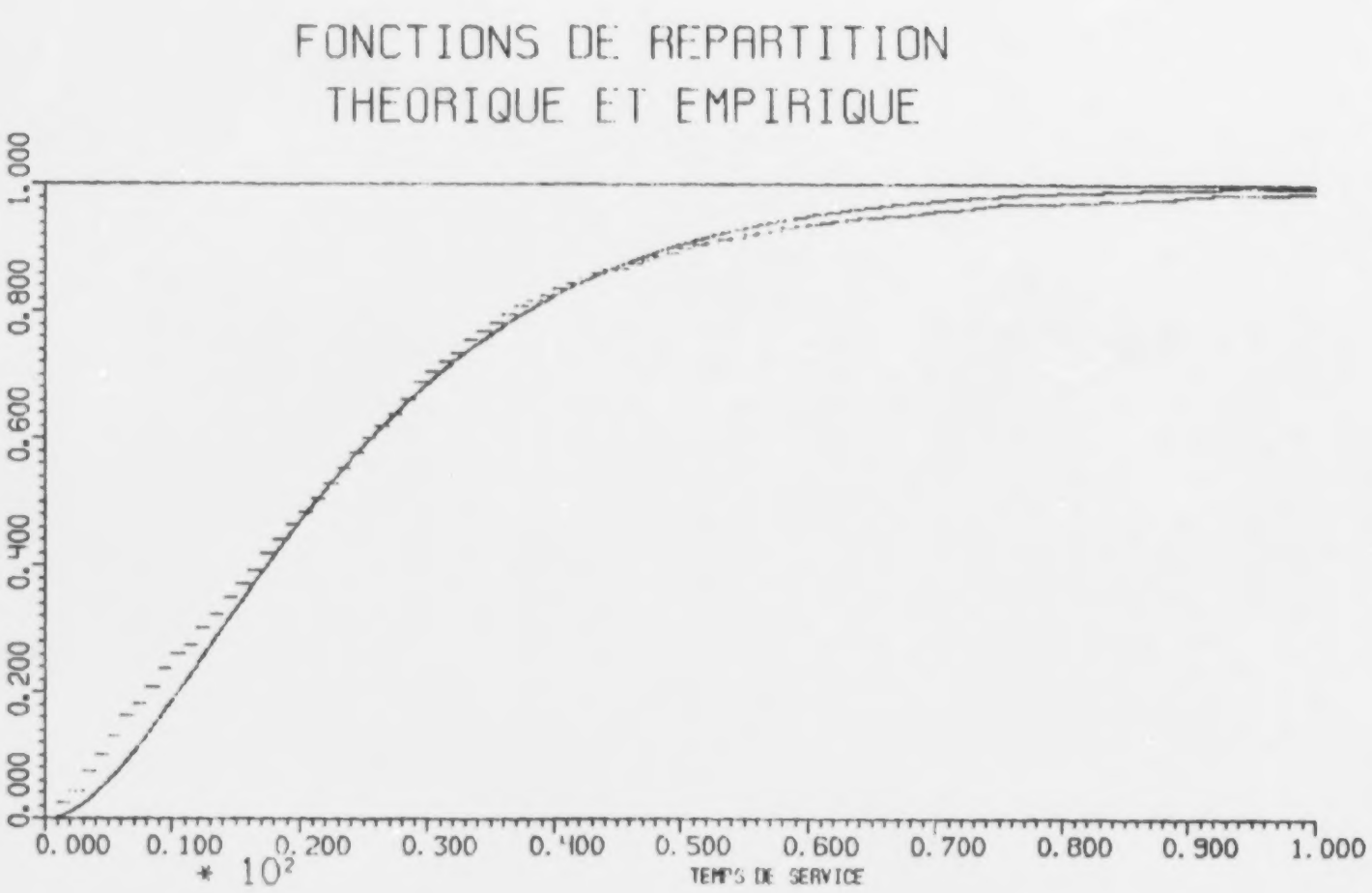
Pour chaque appel, le temps de réponse a été calculé seulement dans le cas où une ambulance était affectée comme ressource première (c'est-à-dire avant l'arrivée d'un médecin). Pour les appels où plusieurs ambulances étaient affectées, on a retenu que le temps de réponse de la première ambulance arrivée sur les lieux. De plus, on a dû écarter les affectations annulées, de même que les cas où au moins une des heures utilisée pour le calcul était erronnée.

En tout, 7356 valeurs ont pu être calculées, soit 76%. Il manque donc des informations, en particulier pour les ambulances affectées comme deuxième ressource. Nous allons quand même étudier cette mesure, mais non sans insister sur le fait que les résultats seront biaisés.

Le tableau 12 donne différentes statistiques sur les temps de réponse par priorité. On y remarque que la répartition des temps est plus étendue pour les appels non prioritaires.

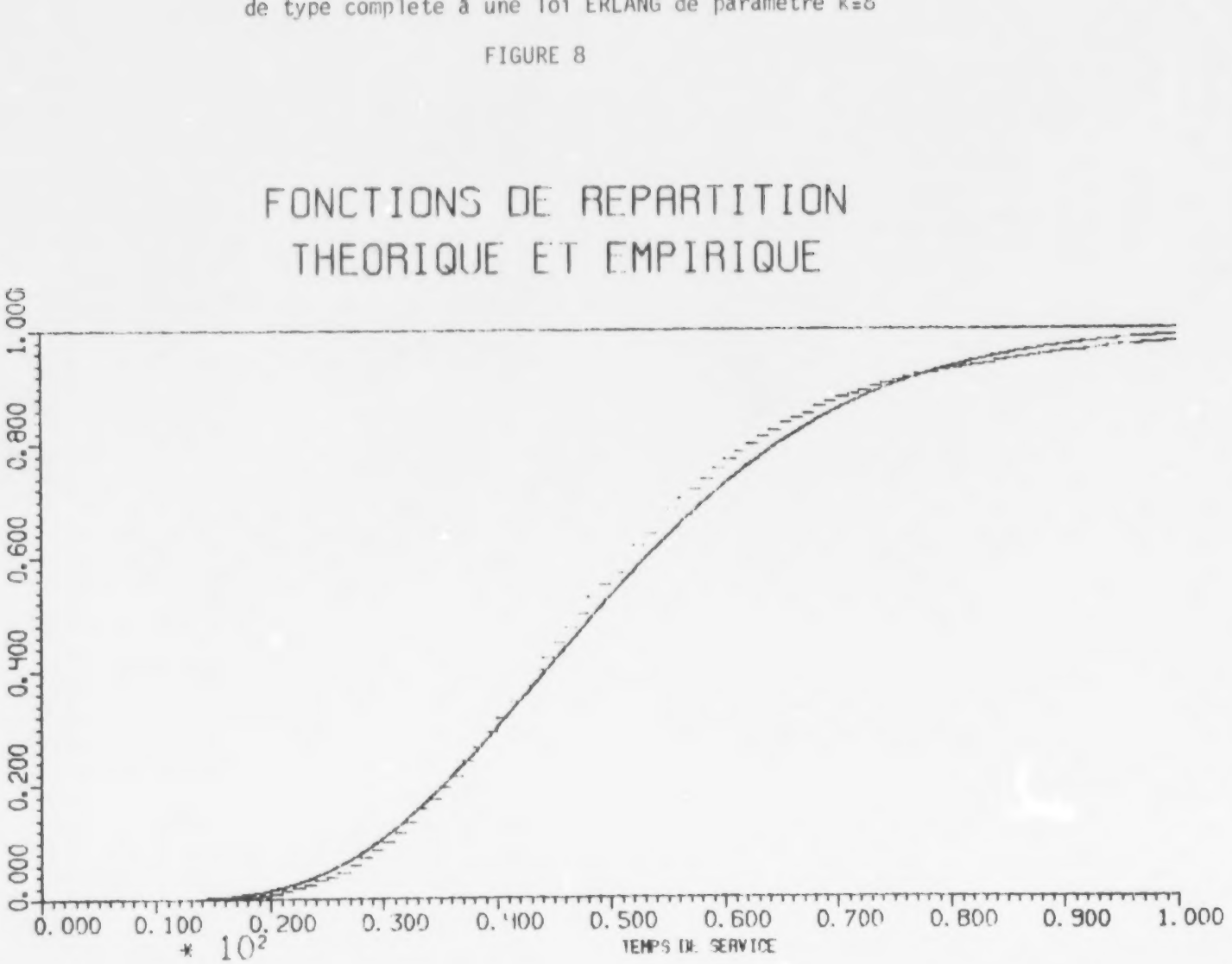
Ajustement des temps de service
de type annulé à une loi ERLANG de paramètre k=2

FIGURE 7



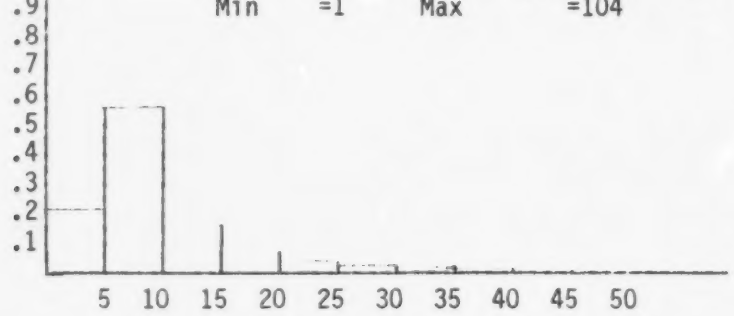
Ajustement des temps de service
de type complete à une loi ERLANG de paramètre k=8

FIGURE 8



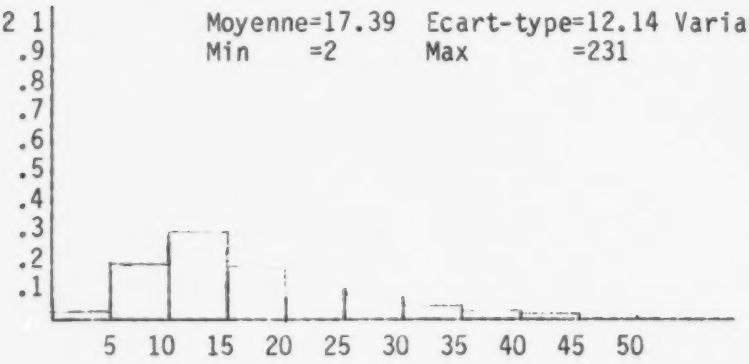
1) Priorité S 1

Moyenne=8.79 Ecart-type=5.59 Variance=31.23
Min =1 Max =104



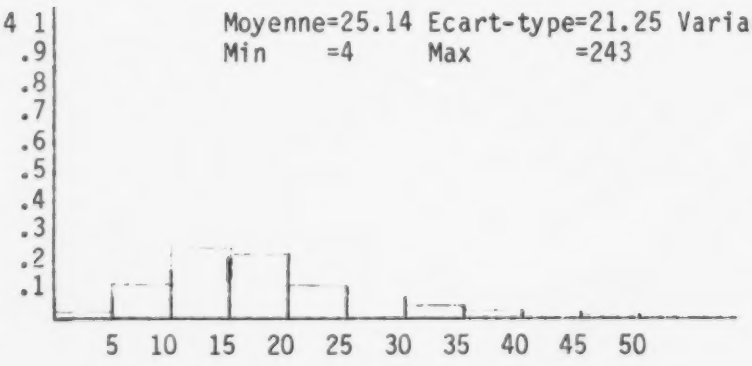
2) Priorité 2 1

Moyenne=17.39 Ecart-type=12.14 Variance=147.47
Min =2 Max =231



4) Priorité 4 1

Moyenne=25.14 Ecart-type=21.25 Variance=451.61
Min =4 Max =243



STATISTIQUES ET HISTOGRAMMES POUR LES TEMPS DE RÉPONSE PAR PRIORITÉ

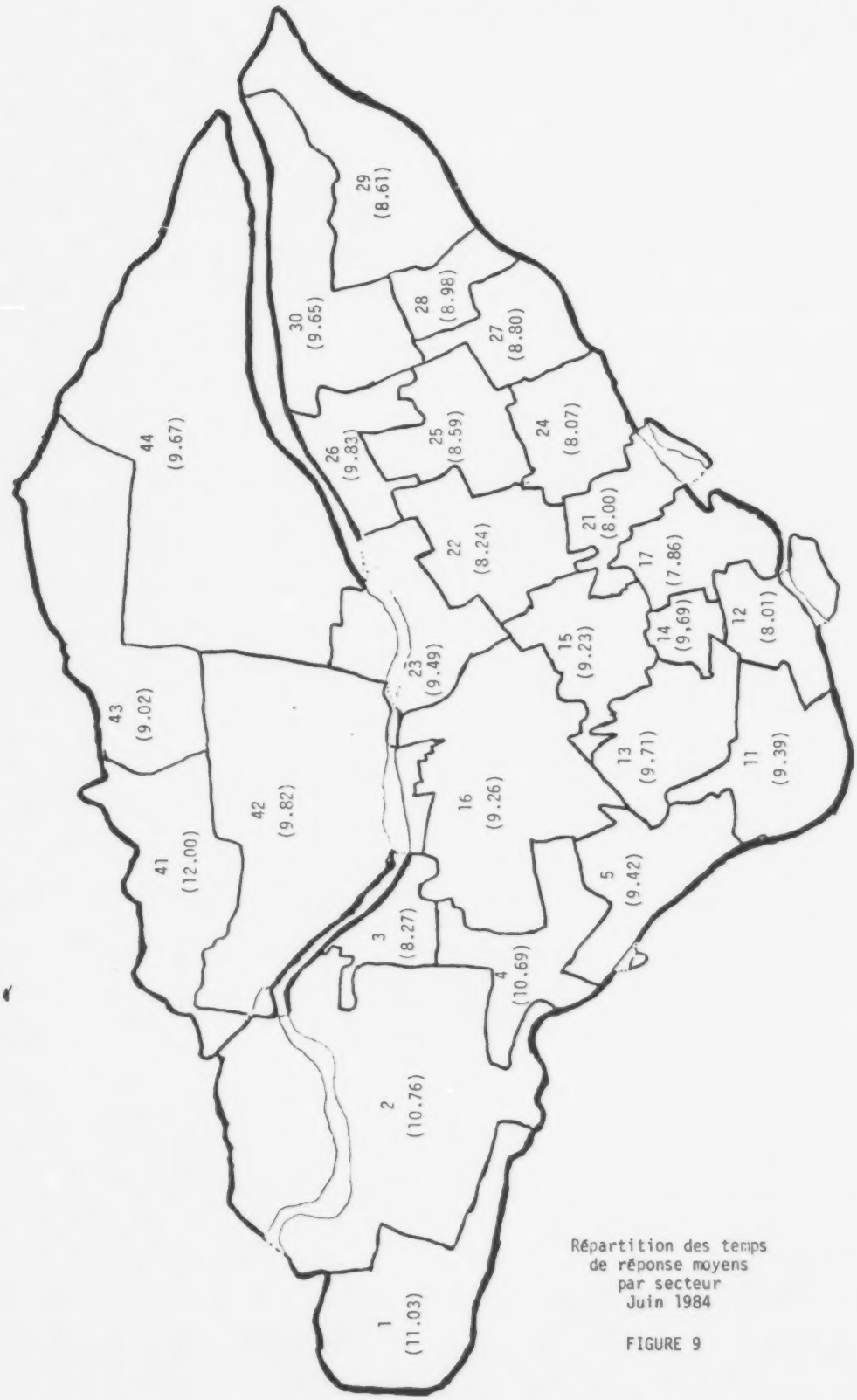
TABLEAU 12

Comme mesure de performance, on s'intéresse particulièrement au pourcentage des appels prioritaires (s) répondus en un temps de dix minutes et moins. Pour les données, ce pourcentage est de 76.7. D'autre part, il est inquiétant de constater qu'un certain nombre d'appels prioritaires (50) soient répondus avec un délai de trente minutes et plus. Pour ces cas, nous sommes allés consulter les archives afin d'obtenir plus d'informations. Ainsi, nous avons remarqué que le délai interne et le temps de déplacement peuvent chacun influencer, substantiellement la longueur du temps de réponse.

Le délai interne peut parfois aller jusqu'à 45 minutes pour un appel de priorité S. L'évaluation médicale du cas, de même que la localisation de l'appel, peuvent nécessiter un temps relativement long. Par exemple, il arrive parfois qu'une urgence soit rapportée par un tiers (opérateur téléphonique, police, ...). Dans ces cas, les informations ne sont pas toujours complètes. Des appels téléphoniques doivent donc être faits afin de compléter les informations. Pour ce qui est du temps de déplacement, il peut prendre des proportions assez grandes (entre 30 et 60 minutes) lorsque l'urgence est localisée en périphérie et/ou lorsque le bénéficiaire est difficile à trouver.

Il est intéressant de visualiser le temps de réponse pour les appels prioritaires (s) en fonction de sa répartition sur le territoire. La figure 9 donne le temps moyen de réponse par région. Les temps les plus grands sont observés dans les régions 1,2,4,26 et 41. Toutes ces dernières sont situées hors du centre-ville.

L'examen du temps de réponse que nous venons de faire est strictement descriptif. Pour ce qui est du reste de notre étude, nous n'allons pas utiliser cette mesure. Nous avons vu que la variabilité de celle-ci dépend en partie d'un certain délai interne entre l'heure de l'appel et l'heure de l'affectation. Notre but n'est pas de minimiser ce délai inter-



Répartition des temps
de réponse moyens
par secteur
Juin 1984

FIGURE 9

ne. Cette question est surtout de nature technique et médicale (prise des informations, identification des lieux et de l'ambulance la plus près, ...). Au chapitre 6, nous considérerons plutôt le temps de déplacement comme mesure des performances.

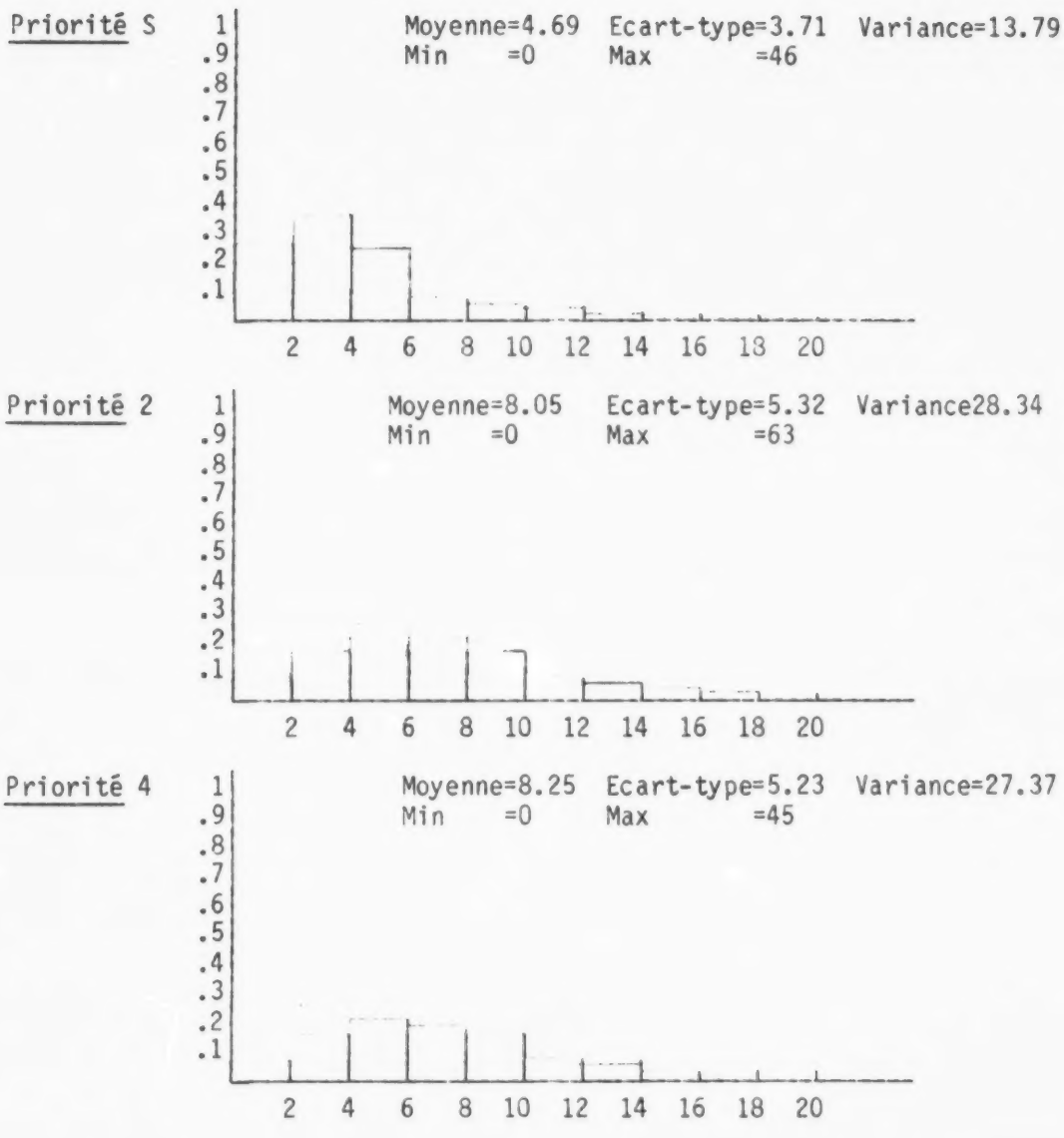
5.5.2 Note sur le temps de déplacement (affectation-arrivée sur les lieux)

Le temps de déplacement a pu être calculé pour 8534 cas (88.5%). Les cas manquants sont répartis ainsi :

heure manquante : 650

ambulance annulée
en cours de route : 452

Le tableau 13 donne différentes statistiques sur le temps de réponse par priorité. À l'examen de celui-ci, on remarque que la répartition des temps de déplacement semble homogène pour les appels de priorité 2 et 4 (moyennes et variances homogènes, similarité des centiles). Le pourcentage des appels prioritaires (S) répondus en un temps inférieur ou égal à 5 minutes est de 72%.



HISTOGRAMMES ET STATISTIQUES SUR LES TEMPS DE DÉPLACEMENT PAR PRIORITÉ

TABLEAU 13

CHAPITRE 6

SIMULATION ET RÉSULTATS

6.1 Rappels des buts

Nous voulons nous prononcer sur la qualité des performances (charges de travail et temps moyen de déplacement global) calculées à partir de la procédure d'approximation du modèle hypercube. Pour un nombre d'ambulances n inférieur ou égal à quinze, Larson a pu calculer exactement les performances en résolvant les 2^n équations d'équilibre. Il a pu ainsi donner une appréciation des performances obtenues par la procédure d'approximation en calculant les erreurs relatives (c'est-à-dire $\left|$ performance exacte - performance approximée $\right| /$ performance exacte) [15].

Dans le cas d'Urgences Santé, le nombre d'ambulances affectées à la couverture, pour les appels de type 4242, est toujours supérieur ou égal à vingt-six. Il est donc impossible d'obtenir les performances exactes, car il faudrait résoudre un système d'au moins $2^{26} \approx 6.7 \times 10^7$ équations linéaires. Afin de se prononcer sur la qualité des approximations, nous aurons besoin d'obtenir des estimations précises des performances exactes. Ces estimations nous seront fournies par un programme de simulation. Nous comparerons les performances calculées analytiquement, à l'aide de la procédure d'approximation, avec celles obtenues à partir du programme de simulation. Ce dernier reproduira exactement le contexte induit par les hypothèses de l'hypercube.

Nous nous intéressons également à mettre le modèle à l'épreuve en nous éloignant des hypothèses qu'il utilise. Nous nous pencherons en particulier sur la robustesse du modèle relativement à l'hypothèse sur les temps de service (issus d'une loi exponentielle de moyenne $1/\mu$ identique

pour tous les serveurs). Il s'agit de voir, pour un contexte où les temps de service ne sont pas exponentiels, si l'erreur commise sur les approximations augmente beaucoup par rapport au premier contexte de l'hypercube (pour lequel les temps de service sont issus d'une loi exponentielle). Encore ici, la simulation nous fournira des estimations précises des performances. Les conditions de simulation seront les mêmes que celles des hypothèses de l'hypercube, à l'exception des temps de service, qui seront issus d'une fonction de répartition empirique.

Note : C'est la procédure d'approximation des charges de travail décrite par Larson ([6], p. 332) qui a été programmée. On suppose que la moyenne des temps de service est la même pour tous les serveurs et que les charges de travail sont équilibrées (homogènes).

6.2 Programme de simulation

6.2.1 Description

L'outil de simulation que nous utiliserons reproduira le contexte des hypothèses de Larson avec une variante sur la fonction de répartition des temps de service. Celle-ci peut être soit exponentielle, soit empirique (nous spécifierons ce terme au paragraphe 6.3). Essentiellement, cet outil de simulation est constitué de trois modules qui sont : le prétraitement, la simulation proprement dite et le module d'analyse des résultats.

Le module de pré-traitement comporte diverses procédures servant à fournir les informations au programme de simulation. Ces informations sont réparties sur quatre fichiers : un fichier d'appels (temps d'arrivée de l'appel et numéro de secteur), un fichier contenant la table des distances entre les secteurs, un fichier contenant la politique de répartition de type "fixed preference" et un fichier contenant les temps de service générés aléatoirement.

Le module central de traitement est utilisé pour simuler l'ensemble des opérations de la flotte d'ambulances, et ce, en reproduisant les conditions des hypothèses de l'hypercube que nous avons décrites au chapitre 4. Entre autres, pour chaque appel lu en entrée, on simule l'envoi d'une seule unité. Celle-ci est déterminée à partir de la politique "fixed preference" qui, dans notre cas, est faite de façon à toujours envoyer d'abord l'unité la plus près. Il n'y a pas de relocalisation des ambulances. Chaque unité retourne instantanément à son poste d'attente lorsqu'elle redevient disponible (c'est-à-dire que le temps de service inclut le retour au garage).

Pour chaque appel, on recueille un ensemble d'informations qui sont écrites sur un fichier de sortie (temps à l'arrivée de l'appel, temps à l'affectation, temps à l'arrivée sur les lieux, temps au retour en disponibilité, numéro du secteur d'où provient l'appel, numéro de l'unité affectée). La conception du programme est inspirée d'un ordinogramme proposé par Larson ([6] chapitre 7).

À partir du fichier de données produit, on peut calculer les statistiques qui nous intéressent (charges de travail et temps moyen de déplacement global, par exemple). Une attention particulière doit être accordée à la construction d'intervalles de confiance à partir des résultats de simulation en raison des observations qui y sont presque toujours corrélées. Le troisième module servira donc à construire des intervalles de confiance pour les charges de travail et pour le temps moyen de déplacement global. Au paragraphe 6.4.1, nous exposerons les méthodes utilisées pour construire ces intervalles.

Les résultats produits ne seront pas le reflet de ce qui se passe dans la réalité. Ils représenteront plutôt une version simplifiée qui correspondra au contexte de l'hypercube. Les trois modules de l'outil de simulation ont été développés au C.R.T.

6.2.2 Base commune pour la simulation et le modèle analytique

Pour comparer les résultats de simulation aux performances approximées, il est nécessaire de porter attention aux paramètres communs à ces deux procédés. Ces paramètres devront être identiques. La base commune pour la simulation et la procédure d'approximation est constituée :

- de la même partition du territoire en 468 secteurs;
- de la même table de distances entre les secteurs;
- du même taux d'appel par secteur;
- du même nombre de véhicules affectés à la couverture et des mêmes postes d'attentes;
- de la même politique de répartition de type "fixed preference";
- du même temps moyen passé auprès du patient, pour tous les serveurs et de la même fonction de répartition de ce temps.

6.3 Description des simulations

6.3.1 Types de simulation

Nous voulons mettre la procédure d'approximation des charges de travail (ρ_i) à l'épreuve dans deux types de situation. Dans le premier, les temps de service seront exponentiels, et dans le deuxième, ils seront empiriques. Les deux cas sont des situations à l'état d'équilibre.

Pour les simulations avec temps de service exponentiels, on simulera exactement les mêmes conditions que celles de l'hypercube. Dans ce type de simulation, on générera des temps de service issus d'une loi exponentielle de moyennes $1/\mu$ identiques pour tous les serveurs. On utilisera la procédure suggérée par Law et Kelton ([7] p. 254). Le paramètre $1/\mu$ a été estimé à partir des données du mois de juin 1984.

Remarque : Par définition, le temps de service est la somme du temps de déplacement et du temps passé auprès du patient. Or, dans la simulation, on génère des temps de service exponentiels alors que le temps de déplacement est déterminé d'une façon indépendante. Il se peut alors que le temps de déplacement soit supérieur au temps de service pour quelques cas, ce qui est alors inconsistant. Pour éviter ce problème, on supposera des vitesses instantanées. On aura donc toujours des temps de déplacement nuls. Cela est évidemment impossible dans un contexte réel, mais ici, ce qui nous intéresse, c'est de vérifier la qualité des ρ_i pour des temps de service exponentiels de moyennes identiques pour tous les serveurs.

Le deuxième type de simulation est nécessaire pour étudier la robustesse du modèle par rapport à l'hypothèse sur les temps de service (exponentiels). Pour chaque appel, on simulera un temps de déplacement non nul. On générera ensuite un temps passé auprès du patient à partir de la fonction de répartition empirique de cette variable. On utilisera la procédure de transformation inverse décrite dans Law et Kelton ([7] p. 261). Le temps de service sera la somme du temps de déplacement et du temps passé auprès du patient.

Remarque : Dans le modèle analytique, les temps de service sont issus d'une fonction de répartition exponentielle de moyenne $1/\mu$ identique pour tous les serveurs. Dans le deuxième type de simulation, les temps de service ne sont pas exponentiels. De plus, le fait de générer des temps de déplacement non nuls peut produire des temps moyens de service différents d'un serveur à l'autre.

6.3.2 Génération des appels

Dans les deux types de simulation, les appels seront générés selon un processus de Poisson d'une manière indépendante de secteur en secteur et à un taux λ_i par secteur i, i=1,2,...,468. On a estimé, pour chaque secteur i, la proportion p_i d'appels provenant du secteur i durant le mois de juin. Pour les différentes situations simulées, on fixera différentes valeur de λ, le taux d'arrivée pour l'ensemble du territoire, et on prendra $\lambda_i = p_i\lambda$. En pratique, on génèrera des temps inter-arrivées issus d'une loi exponentielle de moyenne $1/\lambda_i$ par secteur.

Note : Les proportions p_i sont des proportions moyennes estimées pour l'ensemble de la période du 3 au 23 juin 1984. Ces proportions peuvent varier selon la journée pour un même secteur. Toutefois, nous avons testé l'indépendance de la proportion d'appels par régions (telles que définies au chapitre 5) par rapport aux jours de la semaine. On a obtenu $x_0^2 = 107,93$ avec un nombre degrés de liberté de 150. Puisque $x^2_{c\ 0.05} = 179,30 > x_0^2$, alors on ne rejette pas l'hypothèse nulle au niveau $\alpha = 5\%$. Les proportions d'appels par régions semblent donc peu fluctuer d'un type de journée à l'autre. Ainsi, il n'est pas nécessaire de considérer plusieurs séries d'estimations des p_i pour chaque type de journée.

6.3.3 Situations simulées

Pour chacun des deux types de simulation, on considèrera différentes situations. On fera varier le nombre d'ambulances affectées à la couverture et la charge de travail globale. En tout, on fera douze simulations. Le tableau 6.1 donne les différents paramètres associés à ces simulations.

Temps de service	n	ρ
exponentiel	26	0.8
exponentiel	26	0.4
exponentiel	26	0.2
exponentiel	50	0.8
exponentiel	50	0.4
exponentiel	50	0.2
empirique	26	0.9
empirique	26	0.4
empirique	26	0.2
empirique	50	0.9
empirique	50	0.4
empirique	50	0.2

DESCRIPTION DES DOUZE SIMULATIONS

TABLEAU 6.1

6.4 Analyse des résultats

6.4.1 Construction d'intervalles de confiance pour les mesures

6.4.1.1

Les performances que l'on obtient en simulation sont aléatoires. Dans notre cas par exemple, le temps de déplacement T, pour chaque appel simulé, est une variable aélatoire. Pour avoir une juste appréciation de l'espérance du temps de déplacement, $E(T)=T_0$, inhérent au contexte simulé, on ne doit pas se fier uniquement sur la valeur de la statistique $\bar{T}$ qui est la moyenne des temps de déplacement simulés. Il est préférable de construire un intervalle de confiance pour T_0 qui contiendra T_0 avec une probabilité connue $1-\alpha$, $0 \leq \alpha \leq 1$. On verra plus loin que la précision de cet intervalle est relié à la longueur de la simulation. Ainsi on peut obtenir un intervalle de confiance suffisamment petit pour estimer T_0 précisément.

Les simulations que nous ferons sont des simulations à l'état d'équilibre, par opposition aux simulations terminales. Dans les simulations terminales, les conditions du début et de la fin de l'expérience sont déterminées précisément alors que ce n'est pas le cas pour les simulations à l'état d'équilibre. Dans un processus stochastique, l'état d'équilibre est atteint lorsque les probabilités associées aux états du système ne dépendent plus des conditions initiales. Pour atteindre cet état, on doit passer au début par une période de transition durant laquelle le système se "rode". Cette période varie d'une simulation à l'autre; elle n'est donc pas explicitement déterminée a priori. Pour ce genre de situation, Law et Kelton[7] suggèrent de faire une seule expérience par simulation plutôt que de faire plusieurs expériences indépendantes. L'avantage de faire plusieurs répétitions indépendantes de la même simulation est que l'on obtient, par exemple, plusieurs valeurs indépendantes du temps de déplacement. On peut alors construire un intervalle de confiance par la méthode usuelle (à partir du théorème de la limite centrale et de la loi de Student). Toutefois, dans une situation à l'état d'équilibre, une telle approche n'est pas recommandée en raison du problème qui résulte des conditions initiales différentes d'une expérience à l'autre (qui produisent des périodes de transition de longueurs différentes à chaque fois). Par contre, si on fait une seule expérience pour une même simulation, on aura des observations corrélées. En effet, l'unité disponible préférée pour répondre à un appel dépendra des unités occupées au moment de l'appel. Puisqu'on prend toujours l'unité la plus près, il s'ensuit que la distance parcourue pour répondre à un appel sera corrélée aux distances parcourues pour répondre aux appels précédents. Les distances seront corrélées, et par conséquent, les temps de déplacement le seront aussi. On ne pourra donc pas construire un intervalle de confiance pour T_0 par la méthode usu-

elle (c'est-à-dire en utilisant le T.L.C.). On devra éliminer l'effet de la corrélation en regroupant les observations. C'est la méthode de construction d'intervalles "batch means".

Nous aurons surtout besoin de construire des intervalles de confiance pour les charges de travail ρ_i, i=1,...,n. La construction d'intervalles de confiance multiples pose un problème particulier. Nous allons commencer par exposer brièvement la méthode "batch means" pour une mesure simple avant d'aborder le cas de plusieurs mesures.

Cas d'une mesure simple (Ref. : Law et Kelton[7] p. 295)

Soient T_0, le paramètre à estimer, et T_1, T_2, ..., T_p les p temps de déplacement obtenus durant la simulation après qu'on ait éliminé les premiers appels de la période de transition. Supposons que l'état d'équilibre est atteint. On construit alors m paquets ("batchs") comportant ℓ observations ($m\ell=p$). On définit pour chacun des paquets i, i=1,...,m, le temps moyen de déplacement par

$$\bar{T}_i(\ell) = \frac{\sum_{j=1}^{\ell} T_{(i-1)\ell+j}}{\ell} .$$

On peut alors estimer T_0 sans biais par

$$\bar{\bar{T}}(m,\ell) = \frac{\sum_{i=1}^{m} \bar{T}_{i(\ell)}}{m} .$$

Puisqu'on est dans une situation à l'état d'équilibre, alors $\bar{T}_1(\ell)$, $\bar{T}_2,(\ell),\ldots,\bar{T}_m(\ell)$ sont des variables aléatoires identiquement distribuées de moyenne T_0 et de variance σ^2 inconnue. De plus, si on choisit ℓ assez grand, ces variables seront approximativement indépendantes à cause de l'effet de corrélation qui ne se fera sentir qu'au début et à la fin de chacun des paquets (voir Law et Carson [30]). Si m est assez grand, alors par le théorème de la limite centrale, $\bar{\bar{T}}(m,\ell)$ est approximativement $N(T_0,$

$\frac{\sigma^2}{n}$) et par conséquent, un intervalle de confiance approximatif au niveau $1-\alpha$ pour T_0 est donné par

$$\bar{\bar{T}}(m,\ell) \pm t_{\alpha/2} \sqrt{\frac{S^2\,[\bar{T}_i(\ell)]}{m}}$$

où $$S^2\,[\bar{T}_i(\ell)] = \frac{\sum_{i=1}^{m} (\bar{T}_i(\ell) - \bar{\bar{T}})^2}{m-1}$$

Cas de plusieurs mesures (charges de travail)

On désire estimer par intervalles n charges de travail $\rho_1,\rho_2,\ldots,\rho_n$. Ici, au lieu de déterminer directement le nombre d'observations par paquet, on fixera la longueur des paquets en terme de durée.

Soit T, la durée de la simulation. On construit m paquets, chacun d'une durée Δt, de telle sorte, à ce que $m\Delta t=T$. On veut estimer ρ_i où

$$\rho_i = \lim_{t\to\infty} \frac{\int_0^{\infty} A_i(t)\,dt}{t} \quad \text{et où} \quad A_i = \begin{cases} 1 & \text{si l'unité i est occupée à l'instant t} \\ 0 & \text{sinon} \end{cases}$$

Soient $\rho_{ij}(\Delta t) = \frac{\int_{(j-1)\Delta t}^{j\Delta t} A_i(t)\,dt}{t}$, $j = 1, \ldots m.$

Alors on estime $\rho_{ij}(\Delta t)$ sans biais par la valeur observée de l'intégrale, soit

$$\bar{\rho}_{ij}(\Delta t) = \frac{\text{nombre d'unités de temps où l'unité i est occupée dans le paquet j}}{\Delta t}$$

Les $\rho_{ij}(m,\Delta t)$ sont alors utilisés de la même façon que les $\bar{T}_i(1)$ pour construire un intervalle de confiance pour ρ_i ([7] p. 297).

Pour utiliser cette méthode de construction d'intervalles, on doit porter une attention particulière au nombre de paquets et à leur durée.

Les paquets doivent être assez longs pour être approximativement indépendants et leur nombre doit être suffisamment grand pour que l'on puisse appliquer le théorème de la limite centrale.

L'obtention de plusieurs intervalles de confiance pose un problème pour déterminer le niveau de confiance globale $1-\alpha$. Ainsi, soit I_i l'intervalle de confiance pour ρ_i, $1 \leq i \leq n$, et soit $1-\alpha_i$, le niveau de confiance de l'intervalle I_i. On s'intéresse à la probabilité pour que chacune des charges de travail ρ_i que l'on veut estimer appartienne à son intervalle I_i. C'est le niveau de confiance global pour l'ensemble des paramètres à estimer.

Plus formellement, ce niveau de confiance s'exprime ainsi

$$
\begin{aligned}
1-\alpha &= P\left[\rho_i \in I_i,\ i=1,2,\ldots,n\right] \\
&= P\left[\bigcap_{i=1} (\rho_i \in I_i)\right] \\
&= 1-P\left[\bigcup_{i=1} (\rho_i \notin I_i)\right] \\
&> 1 - \sum_{i=1}^{n} P\,(\rho_i \notin I_i)\Big] \qquad \text{(inégalité de Bonferroni)} \\
&= 1 - \sum_{i=1}^{n} \alpha_i
\end{aligned}
$$

Le véritable niveau de confiance global pour les intervalles sera inconnu, dû au fait que l'on ne peut mesurer exactement la corrélation entre les charges de travail. Tout ce que l'on sait, c'est que le niveau de confiance global sera supérieur ou égal à $1 - \sum_{i=1}^{n} \alpha_i$.

Pour chaque simulation, nous allons construire un ensemble d'intervalles de confiance pour les ρ_i, $i=1,\ldots,n$. Nous adopterons un niveau de confiance global d'au moins 90%. On déterminera la valeur de chacun des α_i par $\alpha_i = \frac{\alpha}{n}$ de sorte que $1-\sum_{i=1}^{n} \alpha_i = .90$.

6.4.2 Déterminer le début et la taille des simulations

Nous n'avons pas fait d'étude pour déterminer le moment où l'état d'équilibre est atteint. Dans toutes les simulations que nous avons fait, nous avons supposé que l'état d'équilibre était atteint à partir de t=500 minutes.

Pour ce qui est de la longueur des simulations, nous avons fait quelques essais qui nous ont permis de déterminer le nombre d'appels nécessaire par simulation en fonction de la précision moyenne voulue sur les intervalles de confiance des ρ_i. (Note : la précision d'un intervalle est définie par sa demi-longueur). Nous avons trouvé qu'avec 30 000 appels, nous aurions une précision moyenne variant entre 0.02 et 0.04. La précision moyenne des intervalles de confiance obtenues pour chacune des douze simulations est toujours strictement inférieure à 0.03.

6.4.3 Résultats

Les tableaux 6.2 et 6.3 montrent l'essentiel des résultats obtenus. Le nombre de charges de travail mal approximées, de même que l'erreur relative commise pour chacune de ces charges, sont les deux principaux résultats qui nous intéressent. Pour les charges de travail ρ_i, l'erreur relative se définit par

$$\frac{\left|\rho_{i \text{ approximée}} - \rho_{i \text{ exacte}}\right|}{\rho_{i \text{ exacte}}}$$

Dans notre expérience, plutôt que d'avoir une valeur fixe pour chaque $\rho_{i \text{ exacte}}$, nous disposerons d'un intervalle de confiance. On calculera l'erreur relative de la manière suivante :

$$\text{erreur relative} = \frac{\left|\rho_{i \text{ approximée}} - m_i\right|}{m_i}$$

où m_i est tel que

$$\left|\rho_{i \text{ approximé}} - m_i\right| = \text{Min}\left\{\left|\rho_{i \text{ approximée}} - a_i\right|, \left|\rho_{i \text{ approximée}} - b_i\right|\right\}$$

et $I_i = [a_i, b_i]$ = intervalle de confiance pour ρ_i.

n	λ	ρ	Variance ρ_i	Précision moyenne I_i	Max$_i$ {Précision I_i}	Nombre de $\rho_i \in I_i$	erreur moyenne	maximale
26	27.87	0.80	0.0023	0.0245	0.0374	0 (0%)	-	-
26	13.93	0.40	0.0044	0.0272	0.0313	3 (11.5%)	0.03	0.06
26	6.97	0.20	0.0017	0.0180	0.0214	6 (23.1%)	0.05	0.09
50	53.59	0.80	0.0105	0.0223	0.0468	15 (30.0%)	0.01	0.07
50	27.79	0.40	0.0153	0.0260	0.0301	18 (36.0%)	0.09	0.50
50	13.40	0.20	0.0050	0.0176	0.0209	25 (50,0%)	0.12	0.71

RÉSULTATS DES SIMULATIONS DE TYPE EXPONENTIEL (charges de travail)

TABLEAU 6.2

n	λ	ρ	Variance ρ_i	Précision moyenne I_i	$\max_i$ Précision I_i	Nombre de $\rho_i \in I_i$	erreur moyenne	maximale
26	23.55	0.95	0.0001	0.0175	0.0322	1 (3.8%)	0.0053	0.0053
26	11.77	0.38	0.0035	0.0226	0.0250	7 (26.9%)	0.0336	0.0606
26	5.89	0.19	0.0018	0.0141	0.0164	11 (42.3%)	0.0657	0.1372
50	45.28	0.90	0.0026	0.0232	0.0630	3 (6.0%)	0.0220	0.0373
50	22.64	0.37	0.0125	0.0217	0.0244	22 (44.0%)	0.0864	0.4129
50	11.32	0.18	0.0039	0.0137	0.0169	27 (54,0%)	0.1064	0.4952

RÉSULTATS DES SIMULATIONS DE TYPE EMPIRIQUE (charges de travail)

TABLEAU 6.3

La précision de chacun des intervalles de confiance I_i est donnée par sa demi-longueur. L'appréciation de la qualité des charges de travail, pour chaque situation, se fait en considérant d'abord le nombre de charges approximées tombant en dehors de leurs intervalles respectifs, et ensuite, par l'erreur relative qui est calculée seulement par les $\rho_i \notin I_i$. Nous avons pu ainsi comptabiliser l'erreur moyenne et l'erreur maximale pour les charges qui ont été mal approximées.

6.4.3.1 Simulations avec temps de service exponentiel (Tableau 6.2)

Pour n=26 et ρ=0.8, toutes les charges de travail approximées tombent dans leur intervalle de confiance respectif. Toutefois, pour ρ=0.4 et ρ=0.2 (toujours avec n=26), le nombre de charges mal approximées est de 3 et 6 alors que l'erreur relative moyenne est de 3 et 5% respectivement.

Lorsque n passe de 26 à 50, on observe une très nette détérioration des résultats. En effet, le pourcentage de charges mal approximées passe

de 0 à 30% pour ρ = 0.8
de 11.5 à 18% pour ρ = 0.4
de 23.1 à 50% pour ρ = 0.2

L'erreur moyenne de ces charges mal approximées passe

de 0 à 1% pour ρ = 0.8
de 3 à 9% pour ρ = 0.4
de 5 à 12% pour ρ = 0.2

De plus, l'erreur maximale prend des proportions inacceptables pour ρ=0.4 (soit 50%) et pour ρ=0.2 (71%) dépassant de loin l'erreur moyenne observée par Larson.

On remarquera que dans la réalité (données de juin 84), on observe un taux moyen d'appels de 13 appels/heure. Ce taux d'appels peut aller jusqu'à 21 à l'heure (données du mois de juin). Compte tenu du nombre de

ressources que l'on affecte, en pratique, pour de tels taux, les charges de travail observées se situent approximativement entre 0.2 et 0.4. Ainsi, c'est dans les situations avec $\rho \leq 0.4$ que le calcul des performances s'avère le plus utile. Or c'est malheureusement dans ces cas, en particulier pour $n=50$, que la procédure d'approximation fournit les résultats les moins bons.

Compte tenu de la seule dimension du problème à Montréal, on ne peut donc affirmer que cette procédure, telle qu'elle est présentement, puisse fournir un calcul fiable des performances du système. Au paragraphe 6.4.3.2 nous suggérons une voie dans laquelle elle pourrait être améliorée.

6.4.3.2 Simulations avec temps de service empiriques

L'examen du tableau 6.4 montre que la situation s'empire une fois de plus lorsque l'on passe d'un temps de service exponentiel à un temps de service empirique. Toutefois, cette détérioration n'est pas très forte, en particulier pour $n=50$. En effet, comparons le nombre de charges mal approximées pour les situations ou ρ est à peu près identique dans les deux types de simulation (c'est-à-dire pour $\rho=0.2$ et $\rho=0.4$).

Pour n = 26, on passe

de 3 (11.5%) à 7 (26.9%) lorsque $\rho \approx 0.2$

et de 6 (23.1%) à 11 (42.3%) lorsque $\rho \approx 0.4$

Pour n = 50, on passe

de 18 (36.0%) à 22 (44.0%) lorsque $\rho \approx 0.2$

de 25 (50.0%) à 27 (54.0%) lorsque $\rho \approx 0.4$

Pour le moment, ces faits nous amènent moins à considérer la robustesse du modèle à l'hypothèse sur les temps de service qu'à renforcer nos doutes quant à la procédure d'approximation elle-même.

n	λ	ρ	T	Intervalle pour T_o	Précision de l'intervalle	Erreur
26	23.55	0.95	18.28	[18.09,18.97]	0.44	-
26	11.77	0.38	5.71	[5.59, 5.72]	0.06	-
26	5.89	0.19	4.98	[4.92, 5.04]	0.06	-
50	45.28	0.90	14.24	[14.92,16,30]	0.60	0.05
50	22.64	0.37	3.74	[3.89, 3.99]	0.05	0.04
50	11.32	0.18	3.09	[3.22, 3.29]	0.04	0.04

ESTIMATION DU TEMPS MOYEN DE DÉPLACEMENT (simulations empiriques)

TABLEAU 6.4

Ainsi, on observe ici le même phénomène que lorsque les temps de service sont exponentiels, c'est-à-dire de moins bonnes approximations des ρ_i pour n=50 par rapport à n=26. De plus, c'est encore pour des taux d'occupation très élevés (0.90 et 0.95), donc irréaliste en pratique, que les ρ_i sont le mieux approximées.

Pour les simulations de type empiriques, nous avons supposé des distances non nulles entre les secteurs (contrairement aux simulations de type exponentiel). Le temps de déplacement a ainsi pu être obtenu. Celui-ci a dans un premier temps été calculé approximativement à partir des charges de travail qui sont elles mêmes des approximations (voir Larson [15]). Dans un deuxième temps, ce même temps de déplacement moyen a été estimé par intervalle à l'aide de la simulation. Le tableau 6.4 donne les résultats obtenus pour les six simulations de type empirique. Le temps moyen approximé tombe dans son intervalle de confiance dans trois situations, soient celles où n=26. Or nous avons vu que c'est pour n=26 que les approximations des ρ_i sont les meilleures. Puisqu'on utilise ces approximations pour calculer le temps moyen de déplacement, il est normal de constater ici de meilleurs résultats pour 26 ambulances. En conséquence, ceci nous amène à chercher une réponse à la question suivante : Pourquoi les situations avec n=26 donne de meilleurs résultats?

6.4.3.3 Une explication : l'homogénéité des charges de travail

Le tableau 6.5 donne la mesure de dispersion des charges de travail pour chaque simulation. Compte tenu d'un même type de simulation et d'un ρ global à peu près identique, on remarque que cette mesure est toujours plus petite pour n=26 par rapport à n=50. Ce qui implique une plus grande homogénéité des ρ_i pour n=26.

Type de simulation	Charge de travail	Nombre d'ambulances	
		26	50
Exponentiel	0.8	0.0023	0.0105
Exponentiel	0.4	0.0044	0.0153
Exponentiel	0.2	0.0017	0.0050
Empirique	0.95	0.0001	0.0026
Empirique	0.4	0.0035	0.0125
Empirique	0.2	0.0018	0.0039

MESURE DE DISPERSION DES ρ_i PAR SIMULATION

TABLEAU 6.5

Note : La mesure de dispersion est la variance usuelle des ρ_i $\dfrac{\sum^{n} (\rho_i - \bar{\rho}_i)^2}{n-1}$

D'autre part, nous avons observé, pour les simulations de type empirique, qu'il n'y a pas beaucoup de détériorations par rapport à celles de type exponentiel. Ce qui semble nous suggérer en particulier que les temps moyens de service différents d'un serveur à l'autre $(\frac{1}{\mu_i} \neq \frac{1}{\mu_j} \quad i \neq j)$ ne soient pas à l'origine des mauvaises approximations observées ici.

Ces considérations nous amènent à penser que les erreurs sont le fruit de la violation de l'hypothèse d'homogénéité des charges de travail. Plus précisément, nous croyons qu'il faille modifier le calcul des $Q(n,\rho,j)$, car c'est à ce niveau qu'intervient l'hypothèse d'homogénéité (voir Larson [6] p. 325).

Globalement, la procédure d'approximation est un processus itératif. Les $Q(n,\rho,j)$ en question y constituent un facteur d'approximation. À chaque itération, ils sont calculés à partir des ρ_i approximés à l'intération précédente. C'est dans la manière de définir ces $Q(n,\rho,j)$ que l'on fait l'hypothèse d'homogénéité sur les ρ_i. Une voie à explorer pour l'amélioration de la procédure consisterait donc à tenter de définir autrement les $Q(n,\rho,j)$ de façon à tenir compte de l'hétérogénéité des ρ_i.

CHAPITRE 7

CONCLUSION

Nous en sommes maintenant à faire le point sur notre étude de l'applicabilité du modèle hypercube. D'abord, nous avons dû vérifier les hypothèses statistiques de base utilisées par le modèle. Pour celle concernant l'arrivée des appels (Poisson), nous avons effectué des tests globaux (pour l'ensemble du territoire) qui n'ont pas véritablement conduit à un rejet de l'hypothèse. Toutefois, il serait intéressant de disposer éventuellement d'observations plus adéquates (c'est-à-dire des temps inter-arrivées non discrets) afin d'effectuer des tests localement (par région) à l'aide d'un test de type Kolmogorov-Smirnov. Pour ce qui est des temps de service, ils ne sont définitivement pas issus d'une loi exponentielle. Cette hypothèse est donc irréaliste, d'où la nécessité d'examiner à fond la robustesse du modèle par rapport à celle-ci.

Outre la vérification des hypothèses de base, l'analyse statistique menée ici nous a permis une description de l'ensemble du système. C'est ainsi que nous avons pu examiner la répartition des appels dans le temps et dans l'espace de même que la répartition des temps de réponse et de déplacement.

À Montréal, la taille du problème est telle que le nombre d'ambulances affectées à la couverture doit en tout temps être supérieur à 15. Le recours à la procédure d'approximation est donc inévitable pour appliquer le modèle hypercube, d'où la nécessité de mettre à l'épreuve cette même procédure.

De son côté, Larson [15] a pu calculer exactement l'erreur relative sur les charges de travail pour un nombre d'ambulances n inférieur à quinze. Il a observé que cette erreur relative ne dépassait que rarement un ou deux pourcent et que les approximations vont en s'améliorant lorsque n croît. Il affirme également que les résultats sont moins bons lorsque les charges de travail sont hétérogènes.

Pour notre part, nous avons testé à l'aide d'un outil de simulation la procédure d'approximation pour $n > 15$. Nous avons estimé l'erreur relative commise sur les charges de travail. Il arrivait dans cinq situations sur six que cette erreur dépassait 2%. Les résultats, dans notre cas, sont moins bons lorsque n est plus grand, quoique nous ne disposions pas d'un échantillon de valeurs de n très important. De plus, c'est pour des taux d'occupation globaux se situant entre 0.2 et 0.4, soient des taux que l'on observe en pratique, que la procédure s'avère la plus inefficace. Nous avons finalement constaté que ces mauvaises approximations étaient reliées à l'hétérogénéité des charges de travail. Il n'est donc absolument pas question d'utiliser la procédure directement. Au pire, on pourrait s'en servir pour discriminer deux solutions.

L'étude de la robustesse du modèle constituait un de nos objectifs. Compte tenu de la faible performance de la procédure d'approximation pour le moment, il n'a pas été possible d'examiner à fond cette question. Toutefois, nous avons constaté que les résultats des simulations montrent une détérioration plus ou moins prononcée lorsque l'on passe d'un temps de service exponentiel à un temps de service issu de la fonction de répartition empirique des données.

Afin de poursuivre le travail ici entrepris, il faudrait d'abord chercher à remédier aux carences de la procédure d'approximation. Comme nous l'avons mentionné à la fin du chapitre 6, nous croyons qu'une voie possible à suivre serait de définir autrement les facteurs d'approximation $Q(n,\rho,j)$ pour tenir compte de l'hétérogénéité des charges de travail.

Avec une procédure fournissant de meilleurs résultats, on pourrait dès lors examiner la question de la robustesse du modèle. L'autre étape consisterait à appliquer et à développer des algorithmes de localisation en contexte non déterministe (comme la procédure de Jarvis [20] pour le problème n-médiane avec congestion).

L'utilisation des solutions que l'on pourrait tirer de tels algorithmes se situe dans un contexte de garage (pas de relocalisation des véhicules). Présentement, ce n'est pas ce contexte qui prévaut à Urgences Santé. Cependant, les algorithmes développés pourraient être utilisés afin d'étudier cette option "garage" et éventuellement appliqués à d'autres types de service comme les pompiers, la police ou les services de réparation à domicile.

REMERCIEMENTS

En tout premier lieu, je tiens à remercier mes deux co-directeurs de recherche, messieurs Jean-Marc Rousseau et Robert Cléroux, tous deux professeurs au département d'Informatique et de Recherche opérationnelle. Je leur suis reconnaissant d'avoir pu bénéficier de leur soutien et de leur expérience.

Je tiens particulièrement à remercier Pierre Trudeau, analyste au Centre de recherche sur les transports pour son étroite collaboration, que j'ai d'ailleurs vivement appréciée.

Pour la partie programmation, un gros merci à Jacques Hodgson et à Pierre Cossette du Centre de recherche sur les transports.

D'Urgences Santé, je suis gré à monsieur Roger Gauthier, à madame Louisette St-Onge, ainsi qu'à tout le personnel de facturation (archives).

Merci à madame Andrée Le Blanc d'avoir si gentiment voulu dactylographier le manuscrit.

Je remercie le Conseil de Recherche en sciences naturelles et en génie (C.R.S.N.G.) pour son appui financier.

Finalement, merci à toi Josée, pour les corrections (!), et surtout, pour ce que tu es.

RÉFÉRENCES

[1] Feller, W. "Introduction to the theory of probability" John Wiley and Sons, 1966 (2 volumes).

[2] Heyman, D., Sobel, M., "Stochastic process in Operations Research", McGraw-Hill, 1982, (2 volumes).

[3] Kleinrock, L., "Queueing Systems", Volume I : Theory, Wiley, New-York, 1975.

[4] Kleinrock, L., "Queueing Systems", Volume II : Computer Applications, Wiley, New-York, 1976.

[5] Larson, R.C., "Urban Police Patrol Analysis", MIT Press, Cambridge, Mass, 1972.

[6] Larson, R.C., Odoni, A.R., "Urban Operations Research", Prentice-Hall, Inc., Englewood Cliffs, N.J., 1981.

[7] Law, M.L., Kelton, D.W., "Simulation modeling and analysis", McGraw-Hill, New-York (1982).

[8] Bratley, P., Fox, B., Schrage, L., "A Guide to simulation", Springer-Verlag, New-York (1983).

[9] Ignall, E.S., Kolesar, P., Swersey, A.J., Walker, W.I., Blum, E.H., Carte, G., Bishop, H., "Improving the deployement of New-York City Fire Compagnies" Interfaces, Vol. 5, No 2, pp. 48-61 (1975).

[10] Carter, C.M., Ignall, E.S., "A Simulation Model of Fire Department Operations : Design and Preliminary Results", IEEE Transactions on system science and Cybernetics, Vol. SSC 6, No.4, October 1970, pp. 282-293.

[11] Ignall, E.S., Kolesar, P., Walker, W.E., "Using Simulation to develop and Validate Analytic Models : Some Case Studies", O.R., Vol. 26, No. 2, p. 237-253.

[12] Kolesar, P., "A Model for Predicting Average Fire Engine Travel Times", O.R., Vol. 23, No. 4, pp 603-613.

[13] Kolesar, P., Walker, W.E., "An Algorithm for Dynamic Relocation of Fire Companies", O.R., Vol. 22, No. 2, pp. 249-274.

[14] Larson, R.C., "A Hypercube Queueing Model for Facility Location and Redistricting in Urban Emergency Services", Computers and Operations Research, 1(1) 67-95 (1974).

REFERENCES (suite)

[15] Larson, R.C., "Approximating the performance of Urban Emergency Service Systems", Operations Research 23(5), 845-868 (1975).

[16] Larson, R.C., "Structural System Models for Locational Decisions : An Example Using the Hypercube Queueing Model",

[17] Handler, G.Y., "Mirchandani, P.B., "Location on Networks Theory and Algorithms", MIT Press, Cambridge, Mass (1979).

[18] Toregas, C. Swain, R., Revelle, C., Bergman, L., "Location of Emergency Service Facilities", O.R. 19, 1366-1373, (1971).

[19] Berman, O., Larson, R.C., "The median problem with congestion", Computers and Operations Research,

[20] Jarvis, J.-P., "Optimization in stochastic service systems with distinguishable servers, Technical Report No. 19-75, Operations Research Center, MIT, Cambridge, Mass. (1975).

[21] Berman, O., "Dynamic Repositionning of indistinguishable Service Units on Transportation Network", Transportation Science, Vol. 15, No. 2, pp. 115-136.

[22] Carter, G.-M., Chaiken, J.M., Ignall, E., "Response Areas for two Emergency Units, Operations Research, 20, 571-594 (1972).

[23] Mallette, C., Lapalme, G., "Guide d'utilisation des Areas Master Files", Centre de recherche sur les transports, Université de Montréal, Pub. #273, novembre 1982.

[24] Roy, S., Chapleau, L., Ferland, J.A., Lapalme, G., Rousseau, J.-M., "The Construction of Routes and Schedules for the Transportation of the Handicaped". Centre de recherche sur les transports, Université de Montréal, Pub. #308, Juin 1983.

[25] Lindgren, B.W., "Statistical Theory", MacMillan Publishing Co. Inc., New-York (1976) (3e édition).

[26] Cramer, H., "Mathematical Methods of Statistics", Princeton, N.J., Princeton University Press, 1946.

[27] Connover, W.S., "Practical Non parametric Statistics", 2e edition, John Wiley and sons, New-York, 1980.

[28] Lilliefors, H.W., "On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown", Journal of the American Statistical Association, 64, pp. 387-389 (1969).

REFERENCES (suite)

[29] Walker, W.E., Chaiken, J.M., Ignall, E.J. et al., "Fire Department Deployment Analysis", A Public policy analysis case study", North-Holland, New-York (1979).

[30] Law, A.M., Carson, J.S., "A Sequential Procedure for Determinating the Lenght of a Steady-State Simulation", Operation Research, 27, 1011-1025 (1979).

[31] Halpern, J., "Finding Minimal Center MEDIAN Convex Combination (Cent-Dian) of a Graph", Management Science, 24, 535-544, (1978).